ENVIRONMENTAL DNA FOR ASSESSINGING SPECIES RICHNESS, GENETIC DIVERSITY, AND SPECIES ABUNDANCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kara J. Andres

December 2022

ENVIRONMENTAL DNA FOR ASSESSINGING SPECIES RICHNESS, GENETIC DIVERSITY, AND SPECIES ABUNDANCE

Kara J. Andres, Ph. D.

Cornell University 2022

ABSTRACT

Environmental DNA (eDNA) approaches, which involve the sampling and analysis of DNA directly from environmental samples, have revolutionized the ability to detect species and monitor biodiversity efficiently and noninvasively. Recent work has demonstrated the potential for eDNA to detect intraspecific genetic variation and conduct population genetic assessments; however, the field of eDNA has thus far been limited to the assessment of short mitochondrial markers that may lack the resolution required for detailed population genetic assessments. In this body of work, I explore the potential for eDNA approaches to assess fish species richness, provide population genetic information, and estimate absolute species abundance. I first combine estimates of species richness using eDNA metabarcoding and multiple capture-based sampling methods to determine the allocation of effort and cost that provides the optimal approach for lake-wide species inventories. Moving beyond species detections, I then review the important considerations and types of analyses that are possible when analyzing intraspecific genetic variation from eDNA samples and conduct a simulation experiment to understand the limitations of estimating species abundance with different genetic markers and levels of genetic variation. I then demonstrate successful amplification of

nuclear microsatellite markers from eDNA samples in a mesocosm experiment, providing accurate estimates of allele frequencies and abundance of an aquatic invasive fish species, the Round Goby (*Neogobius melanostomus*). Next, I use these approaches to conduct a field-based population genetics experiment, showing agreement between tissue-based and eDNA-based estimates of allele frequencies and population genetic parameters. Last, I assess the potential for eDNA approaches to estimate Round Goby abundance in natural environments and compare these estimates to benthic images obtained with an autonomous underwater vehicle (AUV). Collectively, these studies indicate that intraspecific genetic variation from nuclear genetic markers can be detected from eDNA samples in both natural and controlled environments and used to estimate population allele frequencies, genetic parameters, and species abundance. However, careful consideration of the challenges and limitations of these approaches are required for eDNA to be reliably used for population monitoring and assessments.

BIOGRAPHICAL SKETCH

Kara Andres was born in St. Louis, Missouri in 1992 to parents Allan and Ginny. Throughout her childhood, she had the privilege of joining her grandparents on their annual camping trips to the remote wilderness in the Shoshone National Forest in Wyoming. These trips had a tremendous impact on her scientific curiosity and appreciation for the natural world as she was encouraged to explore, ask questions, and document her observations of plants, wildlife, and geology. She was also fortunate to spend time rafting and canoeing the many Missouri streams, where she learned to appreciate the diversity and fragility of these ecosystems. These experiences, coupled with the "don't-come-inside-until-the-streetlights-come-on" parenting style of the time, cemented her love of the outdoors and sharing it with others.

For her undergraduate education, Kara attended Missouri State University (MSU) in Springfield, Missouri. She began her undergraduate career on a pre-med track with an interest in Emergency Medicine and took a position as an undergraduate research assistant studying plant genetics under the guidance of John Heywood. While presenting her findings at the MSU research fair, Kara was fascinated by Day Ligon's keynote talk on alligator snapping turtle conservation in the Ozarks region. For the first time, she imagined herself in a career using science to address some of the major species conservation challenges facing society. That same day, she resolved to pursue a career in scientific research and met with John Havel, who offered her a position as a research assistant at the Trout Lake Research Station in Wisconsin. It was at this iconic limnological field station that Kara developed a particular interest in aquatic science and invasive species.

In 2014, she graduated *magna cum laude* from MSU with an honors B.S. in Biology and Chemistry. Following graduation, Kara took a brief position at Mote Marine Laboratory studying the impacts of ocean acidification on corals before beginning her Master's degree in Biology at Saint Louis University. In Jason Knouft's lab, she delved into stream ecology and completed her thesis on the impacts of modified streamflow regimes on fish morphology. She continued her graduate education in fall 2017 in Cornell's Department of Ecology and Evolutionary Biology.

During her PhD, Kara strived to support her peers and community through commitments to service. She served as a co-leader of Cornell's Diversity Preview Weekend, a program that trains students from underrepresented backgrounds in preparation for graduate school in EEB, and prioritized efforts to implement the program at an institutional level. She also served as a Student Coordinator for the Cornell Atkinson Research Fellows group promoting interdisciplinary collaborative science. Kara was honored to receive the CALS Outstanding Teaching Award in 2019 as well as awards for her research talks at IAGLR, ESA, and the EEB Symposium. She is grateful to have spent her years in graduate school living among the beautiful landscapes of upstate New York and enjoyed much of her free time exploring the gorges, hiking, running, fishing, and kayaking.

In fall 2022, Kara will return to St. Louis as a postdoctoral fellow with the Living Earth Collaborative to study fish microbiomes in some of the same streams she grew up exploring.

ACKNOWLEDGEMENTS

I owe a debt of gratitude to the many people who have offered their support, inspiration, and friendship over the past five years, without whom this dissertation would not be possible. First and foremost, I thank David Lodge, who has been a remarkable advisor. From the day I joined his lab David treated me as a scientific and intellectual equal, providing guidance and support while listening to my ideas and suggestions. He has a special ability to zero in on small details while maintaining a focus on the big picture of any project. I am forever grateful to have learned from someone who is knowledgeable in wide-ranging topics from aquatic science to environmental policy and uses that expertise to conduct impactful science. David was also a tremendous source of uplifting encouragement throughout my time as a graduate student and helped me to remain focused on my overall goals.

I also thank the members of my committee for their scientific brilliance and valuable feedback that has contributed enormously to this work. Jose Andrés largely served a role as an unofficial a co-advisor, frequently meeting with me to discuss research plans, troubleshoot problems, and review results. As I had limited molecular laboratory experience before starting my PhD, Jose patiently shared his expertise in everything from library prep to bioinformatic processing, for which I am tremendously grateful. During my first semester at Cornell, Alex Flecker facilitated an opportunity for me to join a research trip to the Peruvian Amazon, where I developed a fascination of tropical river ecology and learned how to work through the challenges of remote field research. Through our many conversations during this time and since, Alex taught me how to be relentlessly curious and remained enthusiastic about my research even as I pursued work in temperate lakes instead of tropical rivers. I thank Kelly Zamudio for her inspiration and expertise in transforming scientific curiosity to tangible

research projects. My discussions with Kelly would often begin with a bundle of my disjointed ideas and end with a detailed plan for addressing specific questions. Her knowledge of molecular ecology and conservation was invaluable in completing this dissertation.

My work has greatly benefitted from members of the Lodge Lab: Paul Simonin, Paul Czechowski, and Luciana Guimarães de Andrade, as well as undergraduate researchers Maria Haileselassie, Diana Tapia, George Rogalskyj, and Amanda Wong. I also thank the Aquatic Lab Group, of which there are too many people to name, for providing an intellectually stimulating and supportive space to discuss all things aquatic science. I thank my collaborators Suresh Sethi, Peter Esselman, Randy Jackson, Timothy Lambert, Kristy Deiner, Pete McIntyre, and Katie Wagner for their ideas, discussions, and feedback. A heartfelt thanks goes out to the many graduate students and postdoctoral scholars who supported me over the years, especially Erin Larson, Keeley MacNeill, Lily Twining, Rachel Wilkins, Katie Siriani, Coby McDonald, Cinnamon Mittan, Natalie Hofmeister, Elizabeth Lombardi, Amelia Weiss, David Chang Van Oordt, Timothy Lambert, Tram Nguyen, Olivia Graham, Young Ha Suh, Jon Hughes, Kelsey Jensen, Jenny Uehling, Maria Akopyan, Tom Ryan, Jasmine Peters, Jordan Garcia, Ethan Bass, Timothy Salazar, Katie Haviland, Liam Zarri, Megan Barkdull, Noah Brady, Dave Frey, Jenn Houtz, Heath Cook, Marisol Valverde Montellano, Amelia-Juliette Demery, Monique Pipkin, Ethan Duvall, Carl St. John, Kimberly Fitzpatrick, Montana Airey, Paige Muñiz, Sebastian Heilpern, Gemma Clucas, and Nick Ray.

I completed much of the genetic work for this dissertation in the Environmental Genetics Core Facility (EGCF) in our department, which is expertly managed by Steve Bogdanowicz. I am forever indebted to Steve for creating an environment where I was not afraid to ask questions and advice, which likely saved me from more than a few lab disasters.

I am grateful for the research and administrative support of Carol Damm, Jen Holleran, Sally Blinn, Angela Kern, Chad Westmiller, Patty Jordan, Jeremy Searle, Harmony Borchardt-Wier, and Qi Sun. I also thank the staff at Cornell Atkinson who supported me in my role as a Student Coordinator including Carrie Young, Kurt Frijofson, Graham Kerslick, Mark Lawrence, John McKain, Elizabeth Sullivan, Beth Fox, and Christy Goodale.

Many sources of funding made this work possible, including the Department of Defense Strategic Environmental Research and Development Program, National Science Foundation, the Cooperative Institute of Great Lakes Research, Cornell University Graduate School, Sigma Xi, Cornell Atkinson Center for Sustainability, Robert Bouma Fund Research Fund, Andrew W. Mellon Student Research Fund, International Association for Great Lakes Research, Ecological Society of America, Society for Freshwater Science, and the Einaudi Center.

I acknowledge the indigenous lands that are occupied by Cornell University as well as the sites where I completed my dissertation research. I also acknowledge the many Round Gobies whose lives were sacrificed to complete this research. Invasive species attract a villainous reputation, but it is humans that typically create the conditions that lead to ecological and economic harm by these species. I thank the many colleagues whose conversations have helped me rethink my language and relationship with invasive species throughout my time as a graduate student.

Completing graduate school during the COVID-19 pandemic presented many unique challenges. I am endlessly grateful for my “pandemic pod” (Kelsey Jensen, Jenny Uehling, Tom Ryan, and Thom Nelson) that allowed me to safely maintain small amounts of human interaction during a time plagued with uncertainty. I am also grateful for the company of my

Zoom writing crew, Jenny Uehling, Kelsey Jensen, and especially Mabel Baez, for their accountability and encouragement to chip away at this dissertation every day. I owe many fond memories to my volleyball team and the legendary Sea Pigs running group, who motivated me to prioritize movement and pursue personal goals not related to my research. Many close friends have made my time during graduate school particularly special. I especially thank Maria Akopyan for her unfaltering friendship and constant reminder that we are worth more than our scientific output. I am also grateful for the friendship and humor of Kristen Fellows since the day we first met at Mote Marine Laboratory. I thank Stephanie Neitz, Hanna Orf, Olivia Kersting, Michelle Campbell, and Allison Hibbard for continuing to be fantastically supportive and dependable friends.

I would not be here without the encouragement of my parents, Allan and Ginny Andres, who worked endlessly to better understand the idiosyncrasies of academia so they could support me through this time. A particular thanks goes to my sisters Erin and Randi for motivating me to do my best in any endeavor. Many other extended family members provided an endless stream of love and support, and I thank them all for this. I am grateful for my dog Henry who never failed to remind me that long walks and treats are necessary every day. I also thank my soon-to-be in-laws in the Boschert family for always cheering me up with videos of our nieces and nephews. I am thrilled to be moving closer to family where these joyful moments can be shared in person. Last, I thank Joe Boschert for being the most supportive partner imaginable. He enthusiastically encourages my love of science and has been my biggest cheerleader from the day I decided to pursue scientific research. His fascination with critters and nature reminds me daily why I made that decision.

TABLE OF CONTENTS

Abstract iii

Biographical sketch v

Acknowledgements vii

CHAPTER 1 13

Combining sampling gear to optimally inventory species highlights the efficiency of eDNA metabarcoding 13

Abstract 13
Introduction 15
Materials and methods 17
Results 27
Discussion 34
Acknowledgements 42
References 43

CHAPTER 2 53

Detecting and analyzing intraspecific genetic variation with environmental DNA 53

Abstract 53
Introduction 55
Choice of molecular genetic marker 57
Analyzing mitochondrial eDNA 63
Analyzing nuclear eDNA 67
Understanding and accounting for errors 70
Estimating abundance using eDNA 77
Conclusion 91
Acknowledgements 92
References 93

CHAPTER 3 109

Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples 109

Abstract 109
Introduction 111
Materials and methods 115
Results 123
Discussion 131
Acknowledgements 139
References 140

CHAPTER 4 150

Environmental DNA detects population genetic diversity and structure with microsatellite markers 150

Abstract 150
Introduction 151
Materials and methods 154
Results 163
Discussion 173
Acknowledgements 182
References 183

CHAPTER 5 194

Estimating Round Goby abundance using advanced survey methods: a comparison of AUV imaging and eDNA assessments 194

Abstract 194
Introduction 196
Materials and methods 200
Results 212
Discussion 217
Acknowledgements 224
References 225

Appendix 1 239

Appendix 2 248

Appendix 3 251

Appendix 4 258

Appendix 5 271

CHAPTER 1

COMBINING SAMPLING GEAR TO OPTIMALLY INVENTORY SPECIES HIGHLIGHTS THE EFFICIENCY OF EDNA METABARCODING[1]

Abstract

Biodiversity surveys may require the use of multiple types of sampling gear to maximize the efficiency of species detections, yet few studies have investigated how to optimally distribute effort among gear. In this study, we conducted eDNA metabarcoding and capture-based sampling surveys (electrofishing, fyke netting, gillnetting, and seining) to sample fish species richness in a large northern temperate lake. We evaluated the success of the sampling methods individually and in combination to determine the allocation of effort and cost across sampling gear that provides the optimal approach for lake-wide species inventories. We found that eDNA metabarcoding detected more species than any other sampling method, including 11 species that were not detected with any capture-based approach. Optimal gear combination analyses revealed that detected species richness is maximized when most of the effort or budget is allocated to eDNA metabarcoding, with smaller allocations to seining and fyke netting. eDNA metabarcoding and capture sampling gear showed similar patterns of spatial heterogeneity in the fish community across habitat types, with pelagic samples forming a group that was distinct from nearshore samples. Our results indicate that eDNA metabarcoding is a rapid and cost-efficient tool for biodiversity monitoring, and that assessing

[1] This chapter has been accepted for publication as Andres, K. J., Lambert, T.D., Lodge, D. M., Andrés, J., & Jackson, J.R. Combining sampling gear to optimally inventory species highlights the efficiency of eDNA metabarcoding. *In press, Environmental DNA*; it is reprinted here with permission.

the complementarity of multiple sampling types can inform the development of optimal approaches for measuring fish species richness.

Introduction

Comprehensively documenting the status and trends of biodiversity is foundational to ecological research and achieving conservation goals (Margules & Pressey, 2000). However, biodiversity assessments may not fully capture the species richness of an ecosystem, particularly in aquatic ecosystems where heterogeneous and complex habitats can be challenging to sample (Gotelli & Colwell, 2001; Vaux, Whittier, DeCesare, & Kurtenbach, 2000). Such surveys may be limited by time or cost, resulting in incomplete species inventories that are spatially, temporally, or taxonomically biased. Inadequate assessments of the complete species assemblage may hinder efforts to conserve species, particularly those that are rare, elusive, or cryptic (Fleishman, Noss, & Noon, 2006; Gotelli & Colwell, 2011). To efficiently obtain unbiased species richness estimates, sampling survey designs should maximize the number of species detected per unit of sampling effort or cost.

The sampling methods and protocols used to measure species presence and richness are typically chosen based on a compromise between sampling logistics, efficiency, precision, and accuracy (Hughes & Peck, 2008). In aquatic ecosystems, fish species may be sampled using a variety of capture methods in which individuals are caught and identified, most commonly through electrofishing, netting, and trapping (Bonar, Hubert, & Willis, 2009). Non-capture survey methods, including underwater bioacoustics, visual observations, and environmental DNA (eDNA) have also emerged as effective approaches for detecting species in aquatic environments (Deiner et al., 2017; Ellender, Becker, Weyl, & Swartz, 2012; Jerde, Wilson, & Dressler, 2019; Linke et al., 2018; McElroy et al., 2020; Olds et al., 2016; Smart et al., 2016; Thomsen & Willerslev, 2015). eDNA metabarcoding—which uses high-throughput

sequencing to detect the genetic material shed by species—is a highly sensitive approach for surveying species richness that can result in higher detection probabilities (Valentini et al., 2016) and greater species richness estimates (Afzali et al., 2021; Miya et al., 2015; Pont et al., 2018) compared to capture-based survey methods. Given the breadth of aquatic species survey approaches, understanding the limitations and complementarity of each sampling type is an essential component in the development of a comprehensive plan to estimate species richness.

Sampling biases exist for all types of sampling gear with varying degrees of selectivity and specificity, both of which may depend on the biological and physical characteristics of the sampled environment (Pierce et al., 2001). For instance, passive capture sampling gear such as gillnets may overestimate the abundance of large-bodied, fast-swimming species while large mesh sizes may systematically underestimate or exclude small-bodied species (Prchalová et al., 2008; Rudstam, Magnuson, & Tonn, 1984). On the other hand, active capture gear such as electrofishing and beach seining do not efficiently sample benthic species or fast-swimming species that may avoid or escape capture (Mahon, 1980). eDNA metabarcoding is similarly subject to biases in species detections, including those arising from incomplete reference databases favoring certain taxonomic groups (Beng & Corlett, 2020) and primer amplification bias (i.e., the preferential amplification of the sequences of some taxa over others) (Kelly, Shelton, & Gallego, 2019). Because the nature and severity of sampling biases differ among sampling approaches, species inventories in aquatic systems are often completed with survey effort allocated across multiple types of gear to account for different habitat types, species, size selectivity, and efficiency of sampling (Fischer & Quist,

2014). However, it remains unclear how available funds and effort should be distributed among survey approaches to maximize species detections.

In this study, we compare estimates of fish species richness and community composition between multiple capture sampling methods and eDNA metabarcoding. Our objective is to examine the efficiency of the survey approaches independently and in combination, and to use this information to determine how to optimally allocate sampling effort. Following intensive sampling of a fish community with eDNA metabarcoding, electrofishing, fyke netting, gillnetting, and seining, we compare species inventories among gear types and estimate the optimal combination of sampling gear that will maximize the number of species detected across a range of sampling budgets and effort constraints. We also investigate biases in species detections for all sampling types at the lake-wide, and habitat-specific (nearshore vs. pelagic), and site scales to help inform the spatial resolution achieved by different biodiversity survey approaches.

Materials and methods

Study area

Oneida Lake (New York, USA) is a large, shallow mesotrophic lake with a total surface area of 206.7 km^2 and a maximum depth of 16.8 m (mean depth = 6.8 m; Figure 1.1). As the largest lake within the borders of New York State, Oneida Lake supports several water-oriented recreational activities as well as an important sport fishery. Research at the Cornell Biological Field Station (CBFS), supported by Cornell University and the New York Department of Environmental Conservation, has focused intensively on the fisheries and limnology of Oneida Lake since the station was established in 1956. The resulting datasets are

recognized as some of the best long-term data series in aquatic ecology. As of 2016, historical records documented 85 species of fish in Oneida Lake and its tributaries, comprising 52 genera and 22 families (Jackson, 2016). Of these, 77 species have been documented in the lake or its tributaries since 1990.

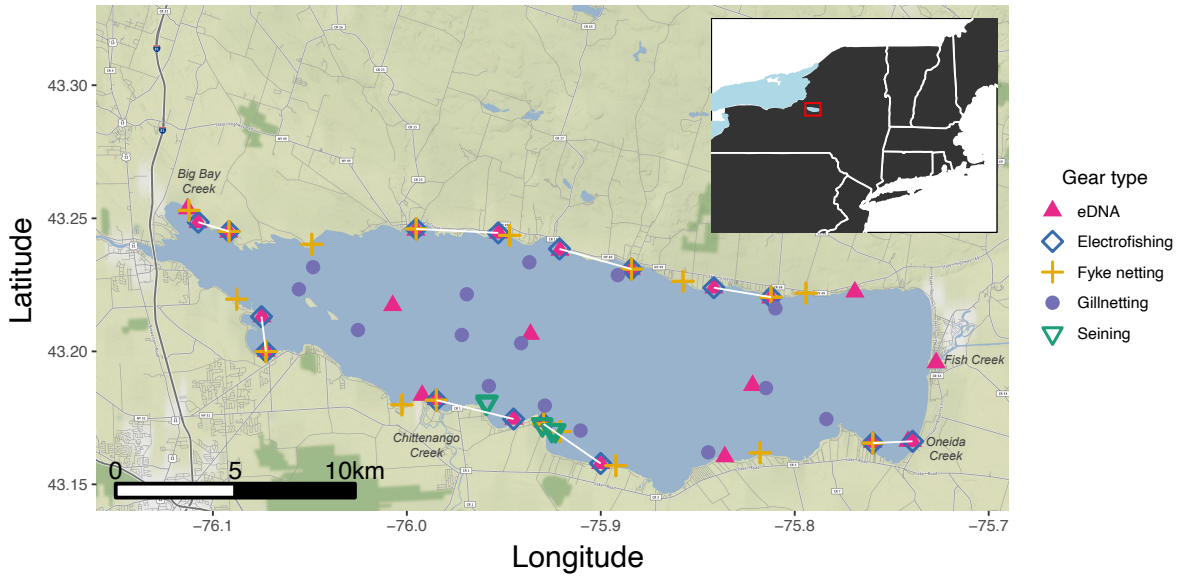

Figure 1.1. Map of eDNA sampling and capture gear survey locations in Oneida Lake, NY, USA. Colors and symbols represent the sampling locations for different gear types. White lines connect the approximate start and end points of each electrofishing transect. Sampled tributary inlets are labeled.

Sampling surveys

Standardized monitoring surveys of Oneida Lake fish populations are conducted annually at the CBFS using a combination of capture methods including electrofishing, gillnetting, fyke netting, and seining. Although all captured species are identified, each

capture survey method is focused on sampling either the nearshore fish community (electrofishing, fyke netting, seining) or the mid-lake, or pelagic, fish community (gillnetting). Electrofishing surveys were conducted from June 12–28, 2017 at each of eight transects distributed around the perimeter of the lake and representing typical shoreline habitats (Figure 1.1). Each electrofishing transect survey involved two 15-minute runs targeting all species for a total of 30 minutes per site. Fyke nets were deployed between September 19–27, 2017 at 18 nearshore sites that were selected to be representative of the proportions of common substrates in nearshore habitats around the entire perimeter of the lake. Each site was sampled for 24 hours with 2 fyke nets, where each net was comprised of a 0.9 m × 1.5 m frame fitted with either 12.7 mm or 5 mm delta knotless mesh. Seine sampling was performed at nine nearshore sites in the southern lake near the CBFS harbor. Each site was sampled monthly from July–September 2017 using a 23 m beach seine with 6.35 mm mesh. Species lists were compiled at the site level for each survey type by combining species identifications for the two survey runs (electrofishing), two nets (fyke netting), and three sampled months (seining).

Pelagic gillnet surveys were used to sample the pelagic fish community at a different standard site each week for 15 consecutive weeks from June 8–September 12, 2017. A multifilament gillnet was deployed on the lake bottom overnight and consisted of four 45.75 m × 1.83 m variable mesh nets sewn together to form a gillnet 183 m in total length. Each of the four nets contained six panels with graded mesh sizes of 38, 51, 64, 76, 89 and 102 mm.

We sampled eDNA in both nearshore and pelagic habitats on June 13, 2017. Nearshore eDNA samples were collected at the start and end of each electrofishing transect (16 locations), as well as at the inlets of Oneida Creek, Fish Creek, Big Bay Creek, and Chittenango Creek (Figure 1.1). Two additional nearshore samples and three pelagic samples

were collected to total 25 eDNA samples. At each sampling site, we collected a water sample from a boat by submerging a sterilized 2-L wide-mouth Nalgene container (Thermo Fisher Scientific, USA) 0.3 m beneath the water's surface. Midway through sampling, a field blank was taken by filling a Nalgene container with distilled water. All samples were stored on ice for up to 12 hours until filtration on the main Cornell University campus (Ithaca, NY) in a laboratory built specifically for eDNA sample analysis consistent with best practices (Goldberg et al., 2016).

Each water sample (1.75 L) was filtered onto 1 μm cellulose-nitrate filters (47 mm diameter; Whatman, GE Healthcare) using a vacuum filtration system. A filtration blank of 1.75 L of distilled water was also processed. Membranes from each sample were immersed in 700 μL Longmire's buffer and stored at -80ºC until DNA extraction using DNeasy Blood and Tissue extraction kits (Qiagen Inc.) with an optimized protocol for eDNA (Spens et al., 2017). If multiple filters were required to filter the full volume of water for a sample, filters were combined in the same microtube prior to extraction.

Library preparation

A NGS library targeting a variable region of the 12S rRNA gene (163–185 bp) was prepared in a two-step process using universal fish primers (MiFish-U-F: GTCGGTAAAACTCGTGCCAGC; MiFish-U-R: CATAGTGGGGTATCTAATCCCAGTTTG; Miya et al., 2015) with either two or three degenerate bases (N) added to increase amplicon heterogeneity. First-stage DNA amplification was conducted in triplicate for each sample and negative control (i.e., field, filtration, extraction, and PCR blanks) in PCR volumes of 20 μL containing 4 μL x5 GoTaq

flexi buffer (Promega, USA), 0.5 μM of each MiFish-U primer (forward and reverse), 0.2 mM of each dNTP, 2 mM $MgCl_2$, 0.15 μL GoTaq HotStart polymerase, 0.4 μg/μL bovine serum albumin (BSA), and 4 μL of eDNA extract as template. The PCR protocol included the following cycling conditions: initial denaturation at 95°C for 2 min, 40 cycles of 95°C for 30 s, 65°C for 60 s, 72°C for 45 s, and a final extension at 72°C for 5 min. PCR products were then pooled by sample and tagged in a second-stage PCR using Illumina Nextera XT tags. The resulting PCR products were purified and size-selected using Agencourt AMPure XP beads (reaction ratio AMPure beads 0.9: PCR product 1; Beckman Coulter Genomics, USA) and the concentration of each library was estimated using the Qubit dsDNA Broad Range Kit and Qubit 2.0 fluorometer (Life Technologies, USA). The libraries were diluted to 18 nM and paired-end sequenced on an Illumina MiSeq sequencing platform with the NextSeq 500 Mid output kit (paired-end 2×150 bp) by Cornell University's Institute of Biotechnology Genomics Facility. Illumina raw sequence data are available from the Dryad Digital Repository.

Sequence data for the 12S gene region are available in public sequence databases (e.g., National Center for Biotechnology Information; NCBI) for most species documented in the Oneida Lake historical datasets. We supplemented the reference database by sequencing tissues from six species collected from Oneida Lake: American Eel (*Anguilla rostrata*), Brook Silverside (*Labidesthes sicculus*), Longnose Gar (*Lepisosteus osseus*), Greater Redhorse (*Moxostoma valenciennesi*), Round Goby (*Neogobius melanostomus*), and Logperch (*Percina caprodes*). DNA was extracted from fin clips sampled from 2–3 individuals per species using manufacturer instructions from a Qiagen DNeasy Blood & Tissue kit and amplified using PCR conditions described above, with DNA template reduced to 2 μL per reaction. PCR

products were purified, quantified, diluted to 34.2 ng per sample, and sequenced at Cornell University's Genomics Facility's Full-Service Sanger Sequencing.

Bioinformatics

Degenerate bases and adaptors were trimmed from demultiplexed paired reads and sequences < 35 bp were removed using Trimmomatic 0.33 (Bolger, Lohse, & Usadel, 2014). Sequences were analyzed using DADA2 1.16 (Callahan et al., 2016), which involved removing forward and reverse primers, trimming reads to 126 bp, and discarding sequences with expected number of errors > 2 (EE, calculated based on Phred scores). Sequences were then denoised using the DADA2 error filtering algorithm, which identifies and discards erroneous sequences based on error models learned from each FASTQ file (i.e., sample). Forward and reverse reads were merged with a minimum overlap of 20 bp and a maximum of one accepted error in the overlap region, and putative chimeras were removed. The resulting dataset contains 12S amplicon sequence variants (ASVs) and the associated read counts for each sample.

Taxonomic assignments for each ASV were obtained using the BLASTn algorithm (Camacho et al., 2009) and the nucleotide and taxonomy (nt/taxdb) databases from NCBI (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2005; Federhen, 2012). We retained the top five target sequence matches for each ASV and assigned species-level taxonomy to ASVs that matched a single species with a sequence identity of ≥ 98%. If multiple species matched the query sequence equally well, ASVs were assigned the lowest common taxonomic rank (genus or family) among the target sequences with equal percent identity. ASVs that could not be classified to the family level were excluded from further analysis.

ASVs with taxonomic assignments were filtered by removing ASVs with read counts < 0.1% of all reads in a sample or with fewer than the average number of non-zero reads summed across all negative controls (23 reads). Non-target species including non-fish and marine species were removed from the dataset.

Optimal gear combinations

Species accumulation curves and optimal gear combinations were derived using information about the sampling effort (work person-hours) and cost ($USD) required to sample a single site for each gear type (Table 1.1, Tables S1.2–S1.4). To provide results that would be applicable beyond our particular study, we made several simplifying choices. We assumed that all sampling took place within fully equipped facilities with streamlined protocols; therefore, we did not allocate effort or cost to capitalization, other startup costs, or methods development. Per-sample effort estimates were based on estimated effort for the field and laboratory work conducted for this study and includes time spent on sample collection, filtration, and laboratory and sequence processing (eDNA), as well as deploying nets and identifying species (capture gear). Because eDNA cost estimates can be highly variable depending on the protocol and number of samples processed, we calculated costs using two approaches: 1) based on a published set of laboratory cost estimates in which all sampling, laboratory, and sequencing is completed by the research institution (Bálint et al., 2018); and 2) based on personal communication with a commercial eDNA metabarcoding service (Smith-Root, Inc.) in which sampling is completed by the research institution and samples are processed by an external company. These cost scenarios are hereafter referred to as "in-house" and "external" costs, respectively. Cost estimates for capture methods included rates

charged by the CBFS for boat usage, which covers boat maintenance, fuel, and depreciation. Both eDNA and capture gear sampling cost estimates also included personnel costs (Table 1.1). A detailed breakdown of estimated sampling effort and costs can be found in Tables S1.2–S1.4.

Table 1.1. Estimated effort (person-hours of work) and cost ($USD) required to obtain one sample for each sampling gear type.

	eDNA (in-house)	eDNA (external)	Electro-fishing	Fyke netting	Gill-netting	Seining
Effort (work hours)	2.0	–	4	3.6	13	2.5
Cost ($USD)	152	284	280	92	320	70

To determine the combination of gear that will maximize species detections at a given amount of effort or cost, we conducted sample-based rarefaction for single- and combined-method surveys. For a dataset with A samples that detects a total of S_{obs} species, the expected number of species observed in a samples is:

$$\mathrm{E}[\mathrm{S}(a)] = \sum_{i=1}^{S_{obs}} \left(1 - \frac{\binom{A - n_i}{a}}{\binom{A}{a}} \right) \quad (1)$$

where n_i is the number of samples in which the i^{th} species is observed (Colwell et al., 2012; Ugland, Gray, & Ellingsen, 2003). For each species, Eqn. 1 calculates the probability of detection as the complement of non-detection. This sample-based rarefaction interpolates the species accumulation curve between $a = 0$ and $a = A$ by assuming samples of a single type are chosen randomly without replacement. We generalize this model to combined-method surveys. Consider a dataset with $k = 1,2, \ldots, m$ different methods, each containing A_k

samples. The expected number of species observed in a survey containing a_1 samples of method 1, a_2 samples of method 2, ..., and a_m samples of method m is:

$$\mathrm{E}[\mathrm{S}(\boldsymbol{a})] = \sum_{i=1}^{S_{obs}} \left(1 - \prod_{k=1}^{m} \frac{\binom{A_k - n_{ik}}{a_k}}{\binom{A_k}{a_k}} \right) \quad (2)$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ defines how the survey allocates samples among the available methods and n_{ik} is the number of samples of method k in which the i^{th} species is observed. In this instance, non-detection requires that a species is not detected in any of the m independent sampling methods.

To find a species accumulation curve for optimal combined-method surveys, we first found the expected number of species observed for all possible combinations of sampling methods up to some maximum total effort or cost, C_{max}. That is, if we let C_k be the cost per sample of method k and $C(\boldsymbol{a}) = \sum_{k=1}^{m} a_k C_k$ be the cost of implementing a survey with sample allocation $\boldsymbol{a}$, we evaluated Eqn. 2 on the set G of all possible values of $\boldsymbol{a}$ such that $C(\boldsymbol{a}) \leq C_{max}$. We then found all non-dominated sample combinations, i.e., all $\boldsymbol{a}$ in G such that there is no $\boldsymbol{x}$ in G where both $C(\boldsymbol{x}) \leq C(\boldsymbol{a})$ and $\mathrm{E}[\mathrm{S}(\boldsymbol{x})] \geq \mathrm{E}[\mathrm{S}(\boldsymbol{a})]$, with at least one of the inequalities being strict. These combinations comprise the Pareto frontier, which describes an effort-based species accumulation curve for surveys that use optimal combinations of sampling methods. Optimal gear combinations are defined as those on the Pareto frontier of all possible combinations of samples from the selected gear, up to a total effort of 100 work hours or total cost of $4,000.

Data analysis

All statistical analyses were performed using R.4.0.4 (R Core Team, 2021) using package *vegan* 2.4-6 (Oksanen et al., 2013). Because sampling approaches may vary in their efficiency in different habitat types, we compared species inventories at lake-wide, habitat-specific, and site scales. For the lake-wide species inventory, we compared species richness observed with eDNA metabarcoding to richness observed by each capture sampling gear alone as well as all types of capture sampling gear combined. We examined the overlap in lake-wide species detections for all combinations of sampling gear and the Oneida Lake historical species records using the *UpSetR* package (Gehlenborg, 2019).

Species inventories at the scale of habitat type involved grouping all sampled sites into nearshore and pelagic habitats and calculating species richness for each sampling method. eDNA samples covered both habitat types, whereas gillnets sampled only pelagic habitats and electrofishing, fyke nets, and seine nets sampled only nearshore habitats. Nonmetric multidimensional scaling (nMDS) based on species presence-absence (*Sørensen* similarity *index*) was used to visualize differences in species composition among habitat types using either eDNA metabarcoding or a combination of all capture gear. For nearshore and pelagic habitats, we calculated species occupancy (i.e., the proportion of sampling sites in which a species is detected) and tested the ordinal association between species occurrence (i.e., presence in a sample) in eDNA samples and capture gear using Kendall's rank correlation coefficient.

Because eDNA samples were collected at the start and end of each of eight electrofishing transects, site-level comparisons are only conducted between eDNA and electrofishing datasets. For this analysis, we combined species identifications from the two

eDNA samples taken per electrofishing transect. We tested for correlations in species occupancy using Spearman's rank-order tests.

Results

Species inventory comparisons

The eDNA sequencing library contained 6,666,951 raw reads, of which 5,735,219 remained after demultiplexing and adaptor trimming. Filtering, merging, and chimera removal in the DADA2 pipeline resulted in further read removal, leaving a total of 3,814,019 reads. Of the 2,982 unique ASVs identified in the eDNA metabarcoding data, 2,310 were assigned taxonomy at the species level, 651 at the genus level, and 20 at the family level. We were able to resolve all but one sequence to at least the family level.

Across all eDNA samples, a total of 41 unique species and 7 unique genus- or family-level taxonomic assignments were identified. Total species richness detected in eDNA samples was higher than electrofishing (total of 31 species), gillnetting (15), fyke netting (27), and seining (27), as well as all capture sampling gear combined (36; Table 1.2; Table S1.1).

Table 1.2. Lake-wide and habitat-specific species richness for the historical Oneida dataset (since 1990), eDNA metabarcoding, and capture sampling gear (electrofishing, gillnetting, fyke netting, and seining). Types of capture gear are displayed both separately and combined, where the number of unique species identified among all capture gears.

Habitat	Historical	eDNA	Capture gear (combined)	Capture gear			
				Electro-fishing	Gill-netting	Fyke netting	Seining
Lake-wide	77	41	36	31	15	27	27
Nearshore	–	40	34	31	–	27	27

Pelagic	–	15	15	–	15	–	–

Eleven species known to be present in Oneida Lake based on the historical dataset were detected with eDNA but none of the capture sampling gear (Figure 1.2; Figure S1.1; Table S1.1). All species detections in capture sampling gear were redundant with at least one other capture sampling gear apart from a single species detected only by seining: *Hybognathus hankinsoni* (Brassy Minnow). Six species were detected in capture sampling gear but not in eDNA metabarcoding: *Cyprinella analostana* (Satinfin Shiner), *H. hankinsoni* (Brassy Minnow), *L. osseus* (Longnose Gar), *Pimephales notatus* (Bluntnose Minnow), *Pimephales promelas* (Fathead Minnow), and *Pomoxis nigromaculatus* (Black Crappie; Figure 1.2; Figure S1.1; Table S1.1). Of these, four could be linked to ASVs that were present in at least one eDNA sample but were either filtered out during data processing (*P. notatus, P. promelas, P. nigromaculatus*) or were ambiguous at the species level during taxonomic assignment (*L. osseus*).

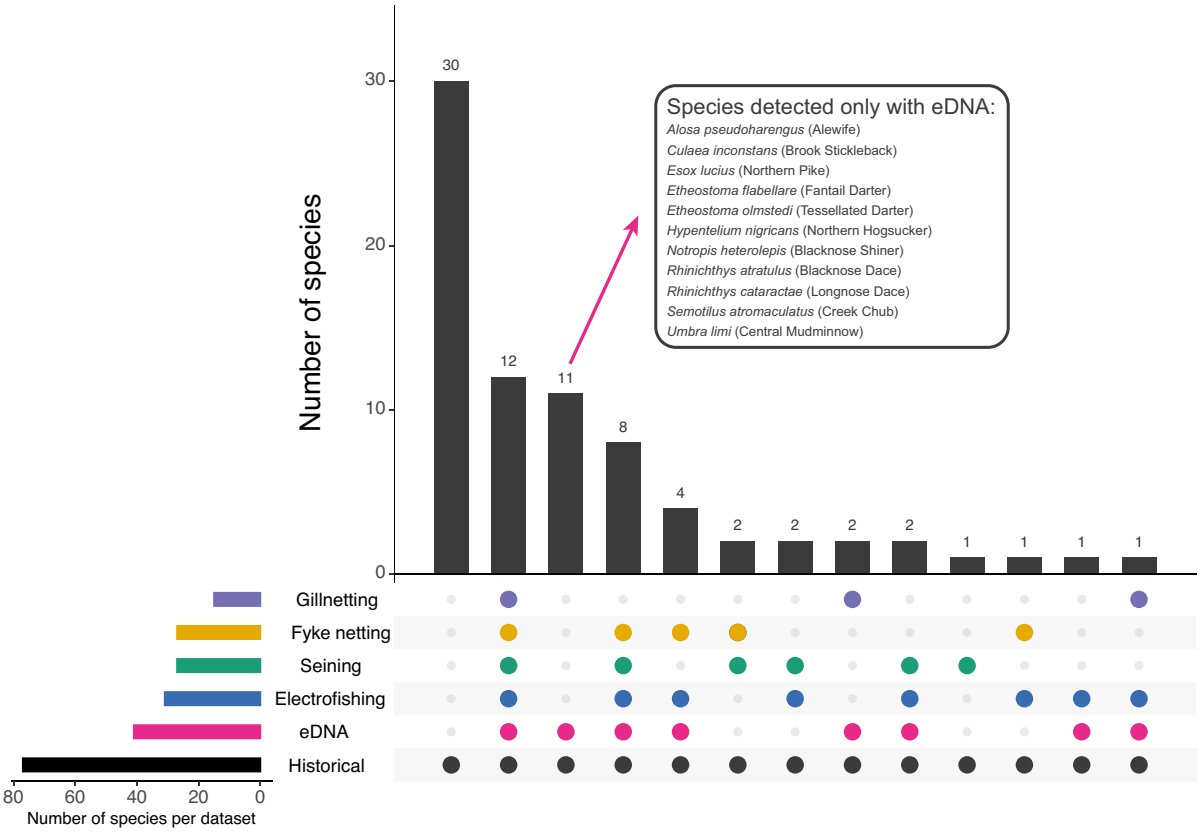

Figure 1.2. Number of detected species uniquely shared by particular combinations of sampling approaches (eDNA metabarcoding, electrofishing, seine, fyke net, and gillnet) and the comprehensive historical species list for Oneida Lake. Dots denote combinations of gear types, and the number of taxa detected by all gear types in a combination (and not by any other gear) is represented by the height of the bar. Combinations of gear types that did not share unique species are not represented in the figure.

Optimal gear combinations

For any given sampling effort, eDNA metabarcoding detected more taxa than capture sampling gear (Figure 1.3A). At low levels of sampling effort (<20 work hours), seining accumulated species more quickly than electrofishing, yet the total number of species detected

using electrofishing exceeded all other types of capture sampling gear when sampling effort increased. Considering optimal combinations of sampling gear, eDNA alone detected nearly as many taxa as all types of gear combined across all sampling efforts (gray line; Figure 1.3A). The stacked bar plot denoting the optimal distributions of sampling effort among gear types shows that a combination of seining and eDNA metabarcoding maximized species detections up to 40 work hours; above this level of total effort, it was beneficial to supplement with small amounts of fyke netting (Figure 1.3D). At 100 hours, the highest total effort that we considered, eDNA is allocated 51% of the effort, while 23% and 26% of the total effort is allocated to seining and fyke netting, respectively. This distribution of effort is equivalent to 25 sites sampled with eDNA, 9 sites with seining, and 7 sites with fyke netting. Electrofishing and gillnetting never detected new species at a rate faster than other gear types could, so they were not allocated work hours at any level of total effort.

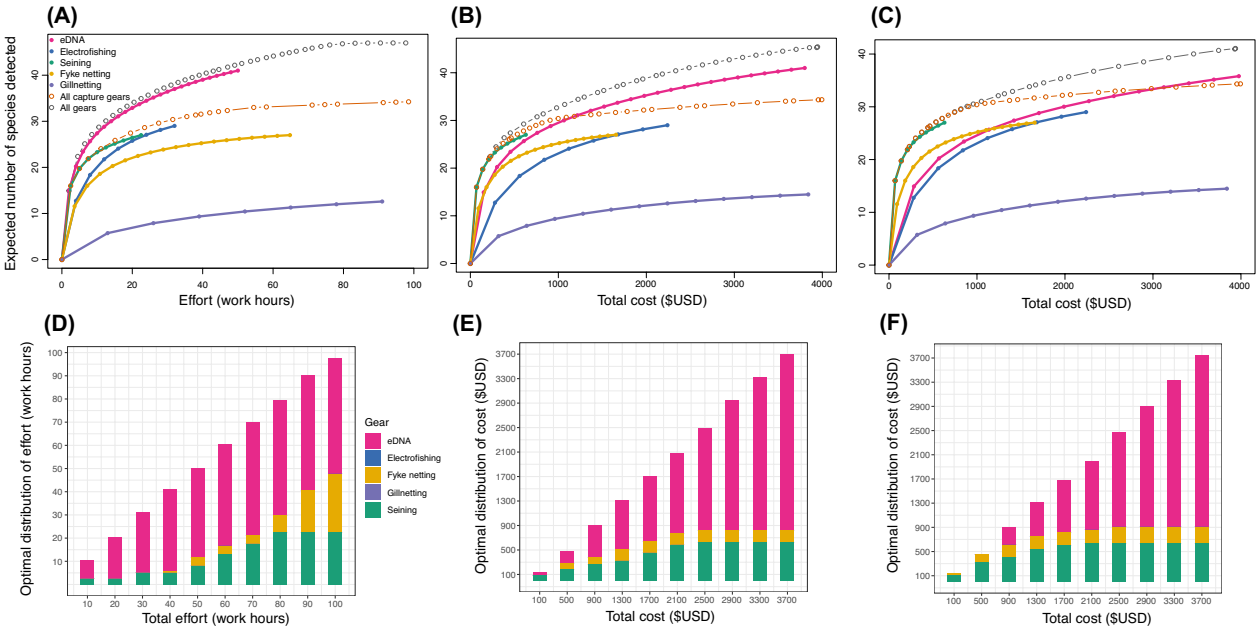

Figure 1.3. Species accumulation curves based on (A) effort (work hours); (B) cost (in-house eDNA); and (C) cost (external eDNA) for each sampling method alone (solid circles) and for optimal combinations of all capture sampling gear (excluding eDNA; hollow orange) and all methods (including eDNA; hollow gray). Stacked bar plots based on (D) effort; (E) cost (in-house eDNA); and (F) cost (external eDNA) indicating the optimal allocation of effort among gear types. For a given total effort, the different colors in each bar indicate how to allocate samples among the five possible types of gear (eDNA, electrofishing, fyke netting, gillnetting, and seining) in a way that maximizes the expected number of species detected.

In the cost-based analyses of gear optimization, seining accumulates species more quickly than eDNA metabarcoding, with eDNA surpassing seining in the total number of species detected if sampling costs exceed $760 (in-house; Figure 1.3B) or $1,420 (external; Figure 1.3C). If eDNA sampling costs are calculated based on in-house estimates, the optimal

distribution of cost is similar to the sampling effort allocations, with most of the budget allocated to eDNA metabarcoding (78%, equivalent to 19 sites) and a smaller proportion allocated to seining (17%, 9 sites) and fyke netting (5%, 2 sites; Figure 1.3E). However, if eDNA sampling costs are calculated based on external estimates, eDNA metabarcoding is not part of the optimal allocation of costs until the budget reaches $900 (Figure 1.3F). Under this scenario, seining and fyke netting dominate the optimal budget allocation when the budget is small, while eDNA consumes 76% (equivalent to 10 sites) of the cost when the maximum budget is reached. The total cost allocated to seining and fyke netting remains low across all sampling budgets, with 17% of the maximum budget allocated to seining (9 sites) and 7% allocated to fyke netting (3 sites). Although species detected with eDNA metabarcoding accumulate more slowly when eDNA is more costly (external scenario; Figure 1.3C) than when eDNA costs are lower (in-house scenario, Figure 1.3B), the sampling effort allocated to eDNA is similar in both scenarios if the maximum budget is consumed (Figure 1.3E–F). Thus, as the sampling budget increases, the relative cost of eDNA does not have as strong of an influence on the optimal distribution of sampling costs among gear types. Neither electrofishing nor gillnetting appeared in the optimal gear combinations for either cost-based optimization.

In all three optimization scenarios, the amount of effort or cost allocated to seining was limited by the number of sampled sites; the amount of seining in a given gear combination cannot exceed the available number of sites sampled by seining in the study. The maximum number of sites sampled by eDNA was also reached in the effort optimization scenario, but not in either cost scenario. Fyke netting did not reach the maximum number of sampled sites at any level of effort or cost.

Habitat-scale and site-level species detections

At the habitat scale, nMDS reveals clear differences in community composition between the species sampled in nearshore versus pelagic habitats in both eDNA metabarcoding (stress = 0.11; Figure 1.4A) and capture-based sampling approaches (stress = 0.16, Figure 1.4B), with no overlap in the convex hulls surrounding the pelagic and nearshore eDNA samples. When comparing species occurrences between eDNA metabarcoding and capture methods, there was a significant correlation in nearshore habitats (Kendall's tau = 0.48, $p < 0.001$) but not in pelagic habitats (Kendall's tau = 0.27, $p = 0.15$; Figure S1.2).

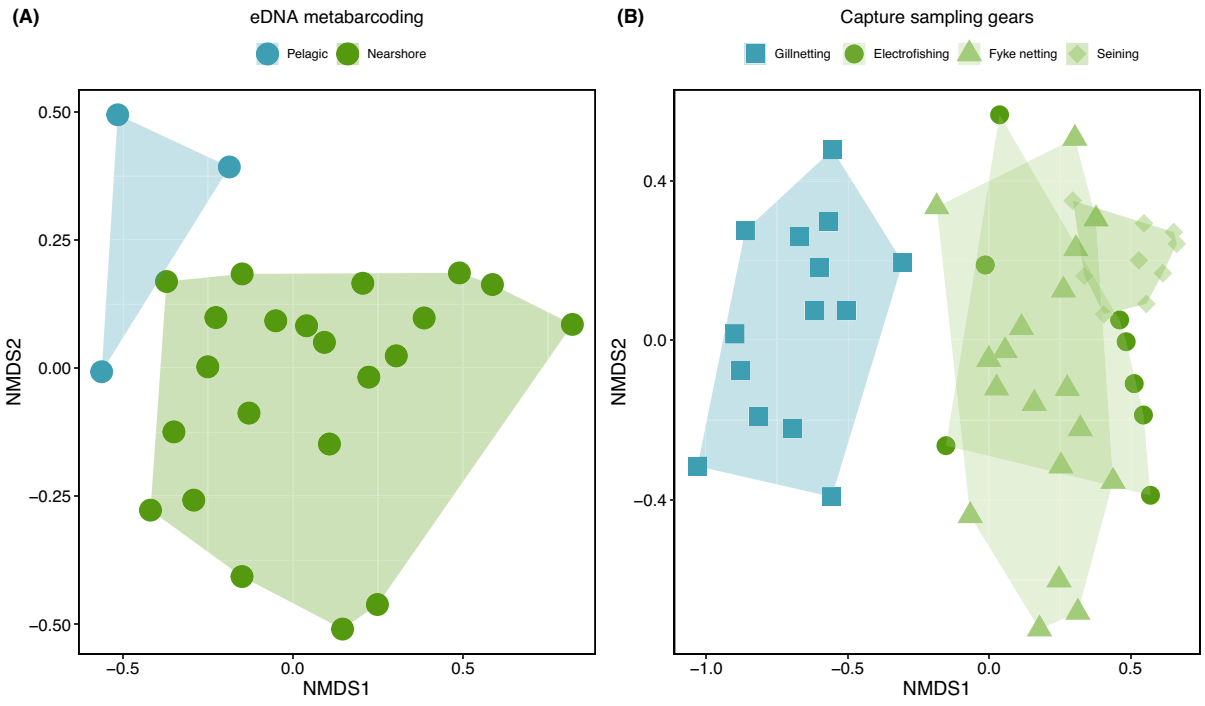

Figure 1.4. Nonmetric multidimensional scaling (nMDS) ordination plots constructed using Sørensen indices of species presence-absence to illustrate habitat differences for eDNA metabarcoding and capture sampling gear. Colors represent habitat types, with (A) eDNA

samples taken in both habitats, and (B) capture sampling methods conducted in pelagic (blue; gillnetting) and nearshore habitats (green; electrofishing, fyke netting, and seining). Convex hulls were constructed to encapsulate all points for each habitat and sampling type.

For the site-level comparisons between eDNA and electrofishing, eDNA detected more species than electrofishing at six of eight transects (Figure S1.3), with higher mean species richness (mean = 19.25, sd = 4.80) than electrofishing (mean = 14.75, sd = 3.84) across all sites, although this difference was not significant (Wilcoxon test, $p = 0.102$). The number of species detected per site from eDNA metabarcoding and electrofishing was not significantly correlated ($r = -0.11$, $p = 0.79$).

Discussion

Freshwater ecosystems face a multitude of ecological stressors and conservation challenges that are contributing to a global reduction in freshwater biodiversity (Reid et al., 2019). Monitoring and preventing further species losses in an era of global change requires strategies to efficiently and rigorously obtain information on the status of species and communities. Over the course of an intensive field sampling effort in a large temperate lake, we assessed the congruency and complementarity of eDNA metabarcoding to capture-based survey methods. We found that eDNA metabarcoding detected more species than any other sampling method, including 11 species historically known to occur in the lake that were not detected with any other survey approach. Species accumulation curves rose more rapidly and saturated at a higher number of species for eDNA metabarcoding than any other sampling method, exceeding the number of species detected with all four types of capture sampling

gear combined. However, a combination of eDNA metabarcoding and capture gear is expected to detect more species than eDNA metabarcoding alone, suggesting that combined-method approaches are required to maximize species detections.

Although many studies suggest that eDNA metabarcoding can be a powerful addition to common capture-based biodiversity monitoring approaches, the complementarity of survey methods is rarely addressed (Beng & Corlett, 2020). When the goal is to maximize detected species richness, our approach finds the optimal allocation of time or cost among sampling gear, accounting for the efficiency of each sampling gear as well as the distinctiveness of the species detected. We found that combined-method species inventories were most efficient when most of the sampling effort or budget was allocated to eDNA metabarcoding and smaller amounts of effort or budget allocated to seining and fyke netting. This general result held regardless of whether eDNA metabarcoding costs were estimated for in-house lab work (lower costs) or with the use of a commercial service (higher costs). Thus, while eDNA metabarcoding is an efficient approach for assessing aquatic species richness in terms of time and cost, the addition of seining and fyke netting is required to optimize species detections. However, other capture-based survey approaches including electrofishing and gillnetting are less efficient or are redundant with other sampling gear and are therefore not included in the optimal method combinations for any scenario.

While eDNA is a promising tool with which to perform lake-wide species inventories, the spatiotemporal resolution of eDNA-based inferences remains less well characterized. Oneida Lake is a shallow, well-mixed lake, and surface currents and seiches may transport eDNA long distances, resulting in species detections at sites where they may not be physically present (Harrison, Sunday, & Rogers, 2019). Furthermore, because DNA fragments may

persist for longer periods of time in sediments than in water (Sakata et al., 2020; Turner, Uy, & Everhart, 2015), remnant DNA from resuspended sediments in the water column may result in the detection of species historically more abundant in the water body. Although we conducted eDNA sampling during a calm weather period to minimize the influence of DNA from resuspended sediments on species detections, the dynamics of eDNA transport, retention, and resuspension in a water body should be considered when assessing species richness using eDNA approaches.

DNA may also be introduced by external sources, including lake tributaries. While all eDNA samples were collected within Oneida Lake, we collected four eDNA samples near tributary inputs (Figure 1.1). However, all but two of the 11 species that we detected only with eDNA sampling have been historically documented within Oneida Lake, with only *Notropis heterolepis* (blacknose shiner) and *Rhinichthys atratulus* (blacknose dace) historically found exclusively in Oneida Lake tributaries (Jackson, 2016). Furthermore, both of these species were detected in a single nearshore eDNA sample that was collected along the northern shore of the lake and not at a tributary inlet. Therefore, although DNA inputs from tributaries may be possible, we did not detect strong evidence of species detections originating from tributaries in this study.

While the exact physical location of species detected with eDNA is challenging to determine, we nevertheless detected habitat-level differences in species composition that were congruent between capture-based surveys and eDNA metabarcoding. For both eDNA and capture survey methods, sites sampled in the pelagic zone were distinct from those sampled in nearshore habitats (Figure 1.4A–B). These results support the growing evidence that eDNA metabarcoding can discriminate habitat-specific eDNA signals in a variety of environments

and sampling approaches (Gehri, Larson, Gruenthal, Sard, & Shi, 2021; Jeunen et al., 2019; Lawson Handley et al., 2019). However, at the site level, there was little association between eDNA-based and electrofishing-based measures of species richness, possibly due to limitations in the ability of eDNA metabarcoding to detect fine-scale heterogeneity in species richness. Thus, while eDNA metabarcoding may be the most efficient approach for performing lake-wide species inventories and detecting broad community differences among habitat types, capture surveys may provide more information about fine-scale (i.e., site-level) species composition and richness. Capture gear may also be beneficial for gathering additional information that may be difficult to assess using eDNA including species abundance, population age structure, and sex ratios, although developments and refinements in eDNA approaches may allow for such metrics to be evaluated in the future (Lacoursière-Roussel, Côté, Leclerc, & Bernatchez, 2016; Yates, Derry, & Cristescu, 2021).

In a single eDNA metabarcoding survey, we detected all but six of the species identified with a combination of four capture sampling gear types throughout an entire sampling season. Upon further inspection, some of these species (*P. notatus, P. promelas, P. nigromaculatus)* were present in the raw eDNA metabarcoding dataset but were removed due to low read counts that were indistinguishable from sampling error. Although we supplemented the genetic database with *L. osseus* (Longnose Gar) 12S sequences, we were nonetheless unable to genetically discriminate between *L. osseus* and *Lepisosteus platostomus* (Shortnose Gar) in the eDNA metabarcoding dataset, indicating that these species may not be distinguishable using the short segment of the 12S gene region targeted in this study. However, because the range of *L. platostomus* does not extend into the sampled region (Fuller, Larson, Makled, & Fusaro, 2022), this species could be excluded if the reference

database was restricted to only species known to occur in Oneida Lake. The remaining two species were infrequently detected in the capture sampling gear, with single specimens of *C. analostana* detected by fyke net and seine surveys, and *H. hankinsoni* collected in a single seine sample.

Maximizing the precision and accuracy of species detections from eDNA metabarcoding may require technical improvements and additional research in both the field and laboratory (Ruppert, Kline, & Rahman, 2019). For instance, some of the differences in species detections between eDNA and capture sampling gear may be due to interactions between DNA and environmental conditions that influence the detection of species (Barnes & Turner, 2015). Because the persistence of eDNA depends on several factors that can vary over time (Barnes et al., 2014; Dejean et al., 2011), there may be considerable measurement stochasticity between sampling events. Repeated sampling through time may prove useful in reducing the associated estimation error and potentially providing a more comprehensive understanding of fish communities at the lake-wide and habitat-level (Bista et al., 2017). Larger eDNA sampling volumes and multiple sampling replicates can reduce sampling stochasticity and allow for assessments of species detection probabilities (Evans, Li, et al., 2017). Furthermore, incorporating information about transport distance and hydrology into eDNA analysis may improve the accuracy of habitat- and site- level species detections (Carraro, Hartikainen, Jokela, Bertuzzo, & Rinaldo, 2018; Harrison et al., 2019). In the laboratory, modifications to the experimental protocol including limiting the number of PCR cycles (Kelly et al., 2019), sequencing samples at greater depth (Alberdi, Aizpurua, Gilbert, & Bohmann, 2018), and analyzing longer or multiple genetic markers (McColl-Gausden et al., 2021; McElroy et al., 2020) can increase taxonomic coverage and reduce taxonomic biases

that arise during PCR and sequencing steps. Although we targeted a single short fragment in this study, a different mitochondrial gene region may have contained sufficient genetic variation to distinguish between closely related species (i.e., *L. osseus* and *L. platostomus*), and multi-marker eDNA metabarcoding methods may be required to maximize species detections (McElroy et al., 2020).

While sampling stochasticity and taxonomic resolution are important considerations in eDNA research, seasonal variation in species composition and abundance must also be acknowledged (Sigsgaard et al., 2017). Although we detected more species in the single day of eDNA sampling than multiple months (June–September) of capture surveys, the detectability of species using all types of sampling gear may vary throughout the year based on seasonal abundance and life history. Therefore, to fully assess the species richness of an ecosystem, temporal changes in species occurrences should be considered when designing sampling surveys. Importantly, because eDNA sampling detects more species with less effort than capture-based sampling gear, eDNA surveys could be conducted multiple times throughout the year, including winter months when deploying capture gear may not be feasible. As noted above, however, the dynamics of eDNA transport, retention, and resuspension may also vary across seasons, and such factors should be considered when designing sampling surveys to maximize species detections.

The fisheries of Oneida Lake have been well-studied for over 60 years with annual monitoring of the whole fish community, enabling the generation of comprehensive species lists, *a priori* expectations of the spatial and temporal dynamics of species presence, and an opportunity to supplement the reference molecular database of genetic markers. Other aquatic ecosystems, such as those with higher levels of biodiversity or poorly described fish faunas,

may require greater investment in ground truthing research before eDNA can be used to reliably characterize species assemblages and abundance (Cilleros et al., 2019). In such cases, the effort and cost expended to build up reference genetic databases and generate species lists may initially reduce the efficiency of eDNA metabarcoding. Ultimately, the efficiency and accuracy of using eDNA metabarcoding is dependent on several factors including availability of reference genetic data, biodiversity of the target assemblage, and the size, complexity, and physicochemical properties of the ecosystem. Regardless of the specific circumstances, our approach for maximizing species detections offers researchers and managers a way to determine how to optimally allocate finite budgets and time across the sampling methods available to them.

eDNA metabarcoding surveys are a powerful and efficient approach for surveying freshwater fish species richness, yet they have not yet been widely adopted by managers as a standardized tool. Questions and concerns are commonly expressed about the sampling effort and costs associated with eDNA surveillance, as well as uncertainty about how eDNA sampling can fit into established monitoring programs (Evans, Shirey, Wieringa, Mahon, & Lamberti, 2017). We demonstrate that eDNA metabarcoding is more efficient than traditional sampling methods for determining species richness, and that eDNA detects habitat-level differences in community composition, in a large, well-studied temperate lake. Some capture-based survey methods, particularly electrofishing and gillnetting, provided little added benefit over other sampling methods in characterizing species richness. However, species detections were maximized with a combination of eDNA metabarcoding, seining, and fyke netting, demonstrating the complementarity of these survey approaches. When designing sampling programs, we recommend that researchers and managers use optimization methods such as

the ones presented in this paper to allocate sampling effort among gear types in accordance with the study objectives, available resources, gear efficiency, and desired outcomes.

Acknowledgements

We thank Lindsay Schaffner, Anthony VanDeValk, Tom Brooking, and Brian Young for assistance with data collection, Lars Rudstam for providing cost estimates for sampling gear, and the New York State Department of Environmental Conservation for providing tissue samples. We also acknowledge the Biotechnology Resource Center Genomics Facility at the Cornell Institute of Biotechnology. Funding for this project was supported by the Cornell Atkinson Center for Sustainability, a Cooperative Institute for Great Lakes Research Graduate Research Fellowship, a Presidential Life Science Fellowship (Cornell University), a National Science Foundation Graduate Research Fellowship (DGE-1650441), and the New York State Department of Environmental Conservation (Contract #C009328). All sampling was conducted under approved Cornell Animal Care and Use Protocol 2006-0088.

References

Afzali, S. F., Bourdages, H., Laporte, M., Mérot, C., Normandeau, E., Audet, C., & Bernatchez, L. (2021). Comparing environmental metabarcoding and trawling survey of demersal fish communities in the Gulf of St. Lawrence, Canada. *Environmental DNA, 3*(1), 22-42.

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution, 9*(1), 134-147.

Bálint, M., Nowak, C., Márton, O., Pauls, S. U., Wittwer, C., Aramayo, J. L., . . . Jansen, M. (2018). Accuracy, limitations and cost efficiency of eDNA-based community survey in tropical frogs. *Molecular Ecology Resources, 18*(6), 1415-1426.

Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics, 17*(1), 1-17. doi:10.1007/s10592-015-0775-4

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology, 48*(3), 1819-1827.

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation, 29*(7), 2089-2121.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research, 33*(suppl_1), D34-D38.

Bista, I., Carvalho, G. R., Walsh, K., Seymour, M., Hajibabaei, M., Lallias, D., . . . Creer, S. (2017). Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nature Communications, 8*(1), 1-11.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bonar, S. A., Hubert, W. A., & Willis, D. W. (2009). Standard methods for sampling North American freshwater fishes. *American Fisheries Society.*

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods, 13*(7), 581-583. doi:10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics, 10*(1), 1-9.

Carraro, L., Hartikainen, H., Jokela, J., Bertuzzo, E., & Rinaldo, A. (2018). Estimating species distribution and abundance in river networks using environmental DNA. *Proceedings of the National Academy of Sciences, 115*(46), 11724-11729.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., . . . Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity

environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources, 19*(1), 27-46.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology, 5*(1), 3-21.

Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., . . . Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895. doi:10.1111/mec.14350

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS One, 6*(8), e23398.

Ellender, B. R., Becker, A., Weyl, O. L., & Swartz, E. R. (2012). Underwater video analysis as a non-destructive alternative to electrofishing for sampling imperilled headwater stream fishes. *Aquatic Conservation: Marine and Freshwater Ecosystems, 22*(1), 58-65.

Evans, N. T., Li, Y., Renshaw, M. A., Olds, B. P., Deiner, K., Turner, C. R., . . . Pfrender, M. E. (2017). Fish community assessment with eDNA metabarcoding: effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences, 74*(9), 1362-1374.

Evans, N. T., Shirey, P. D., Wieringa, J. G., Mahon, A. R., & Lamberti, G. A. (2017). Comparative cost and effort of fish distribution detection via environmental DNA analysis and electrofishing. *Fisheries, 42*(2), 90-99.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research, 40*(D1), D136-D143.

Fischer, J. R., & Quist, M. C. (2014). Characterizing lentic freshwater fish assemblages using multiple sampling methods. *Environmental Monitoring and Assessment, 186*(7), 4461-4474.

Fleishman, E., Noss, R. F., & Noon, B. R. (2006). Utility and limitations of species richness metrics for conservation planning. *Ecological Indicators, 6*(3), 543-553.

Fuller, P., Larson, J., Makled, T. H., & Fusaro, A. (2022). *Lepisosteus platostomus* Rafinesque, 1820: U.S. Geological Survey, Nonindigenous Aquatic Species Database. Retrieved from https://nas.er.usgs.gov/queries/FactSheet.aspx?SpeciesID=757

Gehlenborg, N. (2019). UpSetR: a more scalable alternative to Venn and Euler diagrams for visualizing intersecting sets. *R package version, 1*(0).

Gehri, R. R., Larson, W. A., Gruenthal, K., Sard, N. M., & Shi, Y. (2021). eDNA metabarcoding outperforms traditional fisheries sampling and reveals fine-scale heterogeneity in a temperate freshwater lake. *Environmental DNA, 3*(5), 912-929.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., . . . Taberlet, P. (2016). Critical considerations for the application of environmental DNA

methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307.

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters, 4*(4), 379-391.

Gotelli, N. J., & Colwell, R. K. (2011). Estimating species richness. In: *Biological diversity: frontiers in measurement and assessment, 12*(39-54), 35.

Harrison, J. B., Sunday, J. M., & Rogers, S. M. (2019). Predicting the fate of eDNA in the environment and implications for studying biodiversity. *Proceedings of the Royal Society B, 286*(1915), 20191409.

Hughes, R. M., & Peck, D. V. (2008). Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society, 27*(4), 837-859.

Jackson, J. R. (2016). Oneida Lake BC (Before Cornell). In L. G. Rudstam, E. L. Mills, J. R. Jackson, & D.J. Stewart (Eds.), *Oneida Lake: long-term dynamics of a managed ecosystem and its fishery*. (p. 31–52). Bethesda, Maryland, USA: American Fisheries Society.

Jerde, C. L., Wilson, E. A., & Dressler, T. L. (2019). Measuring global fish species richness with eDNA metabarcoding. *Molecular Ecology Resources, 19*(1), 19-22.

Jeunen, G. J., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., . . . Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong

discrimination among diverse marine habitats connected by water movement. *Molecular Ecology Resources, 19*(2), 426-438.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Scientific Reports, 9*(1), 12133. doi:10.1038/s41598-019-48546-x

Lacoursière-Roussel, A., Côté, G., Leclerc, V., & Bernatchez, L. (2016). Quantifying relative fish abundance with eDNA: a promising tool for fisheries management. *Journal of Applied Ecology, 53*(4), 1148-1157.

Lawson Handley, L., Read, D. S., Winfield, I. J., Kimbell, H., Johnson, H., Li, J., . . . Hänfling, B. (2019). Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environmental DNA, 1*(1), 26-39.

Linke, S., Gifford, T., Desjonquères, C., Tonolla, D., Aubin, T., Barclay, L., . . . Sueur, J. (2018). Freshwater ecoacoustics as a tool for continuous ecosystem monitoring. *Frontiers in Ecology and the Environment, 16*(4), 231-238.

Mahon, R. (1980). Accuracy of catch-effort methods for estimating fish density and biomass in streams. *Environmental Biology of Fishes, 5*(4), 343-363.

Margules, C. R., & Pressey, R. L. (2000). Systematic conservation planning. *Nature, 405*(6783), 243-253.

McColl-Gausden, E. F., Weeks, A. R., Coleman, R. A., Robinson, K. L., Song, S., Raadik, T. A., & Tingley, R. (2021). Multispecies models reveal that eDNA metabarcoding is

more sensitive than backpack electrofishing for conducting fish surveys in freshwater streams. *Molecular Ecology, 30*(13), 3111-3126.

McElroy, M. E., Dressler, T. L., Titcomb, G. C., Wilson, E. A., Deiner, K., Dudley, T. L., . . . Jerde, C. L. (2020). Calibrating environmental DNA metabarcoding to conventional surveys for measuring fish species richness. *Frontiers in Ecology and Evolution, 8*, 276.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., . . . Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science, 2*(7), 150088. doi:10.1098/rsos.150088

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., . . . Oksanen, M. J. (2013). Package 'vegan'. *Community ecology package, version*, *2*(9), 1-295.

Olds, B. P., Jerde, C. L., Renshaw, M. A., Li, Y., Evans, N. T., Turner, C. R., . . . Lamberti, G. A. (2016). Estimating species richness using environmental DNA. *Ecology and Evolution, 6*(12), 4214-4226. doi:10.1002/ece3.2186

Pierce, C. L., Corcoran, A. M., Gronbach, A. N., Hsia, S., Mullarkey, B. J., & Schwartzhoff, A. J. (2001). Influence of diel period on electrofishing and beach seining assessments of littoral fish assemblages. *North American Journal of Fisheries Management, 21*(4), 918-926.

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., . . . Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports, 8*(1), 1-13.

Prchalová, M., Kubečka, J., Říha, M., Litvín, R., Čech, M., Frouzová, J., . . . Vašek, M. (2008). Overestimation of percid fishes (Percidae) in gillnet sampling. *Fisheries Research, 91*(1), 79-87.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T., . . . Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews, 94*(3), 849-873.

Rudstam, L. G., Magnuson, J. J., & Tonn, W. M. (1984). Size selectivity of passive fishing gear: a correction for encounter probability applied to gill nets. *Canadian Journal of Fisheries and Aquatic Sciences, 41*(8), 1252-1255.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation, 17*, e00547.

Sakata, M. K., Yamamoto, S., Gotoh, R. O., Miya, M., Yamanaka, H., & Minamoto, T. (2020). Sedimentary eDNA provides different information on timescale and fish species composition compared with aqueous eDNA. *Environmental DNA*, *2*(4), 505-518.

Sigsgaard, E. E., Nielsen, I. B., Carl, H., Krag, M. A., Knudsen, S. W., Xing, Y., ... & Thomsen, P. F. (2017). Seawater environmental DNA reflects seasonality of a coastal fish community. *Marine Biology*, *164*(6), 1-15.

Smart, A. S., Weeks, A. R., van Rooyen, A. R., Moore, A., McCarthy, M. A., & Tingley, R. (2016). Assessing the cost-efficiency of environmental DNA sampling. *Methods in Ecology and Evolution, 7*(11), 1291-1298.

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S. T., . . . Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution, 8*(5), 635-645. doi:10.1111/2041-210x.12683

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation, 183*, 4-18. doi:10.1016/j.biocon.2014.11.019

Turner, C. R., Uy, K. L., & Everhart, R. C. (2015). Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation*, *183*, 93-102.

Ugland, K. I., Gray, J. S., & Ellingsen, K. E. (2003). The species–accumulation curve and estimation of species richness. *Journal of Animal Ecology, 72*(5), 888-897.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., . . . Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Vaux, P. D., Whittier, T. R., DeCesare, G., & Kurtenbach, J. P. (2000). Evaluation of a backpack electrofishing unit for multiple lake surveys of fish assemblage structure. *North American Journal of Fisheries Management, 20*(1), 168-179.

Yates, M. C., Derry, A. M., & Cristescu, M. E. (2021). Environmental RNA: a revolution in ecological resolution? *Trends in Ecology & Evolution, 36*(7), 601-609.

CHAPTER 2

DETECTING AND ANALYZING INTRASPECIFIC GENETIC VARIATION WITH ENVIRONMENTAL DNA

Abstract

Advancements in environmental DNA (eDNA) approaches have allowed for rapid and efficient species detections in diverse environments. Although most eDNA research is focused on leveraging genetic diversity to identify taxa, some recent studies have explored the potential for these approaches to detect within-species genetic variation, allowing for population genetic assessments and abundance estimates from environmental samples. Here, we discuss the opportunities and challenges in assessing intraspecific genetic variation in eDNA studies. First, we discuss the different types of molecular genetic markers that may be used for eDNA research and summarize the factors researchers must consider when selecting markers to study intraspecific genetic variation. We then discuss the types of population genetic analyses that are possible using different marker types, and how these approaches differ from common tissue-based analyses. We highlight the importance of data cleaning, as the removal of sequence errors is paramount in the analysis of intraspecific diversity from eDNA. Finally, we outline how species-specific abundance estimates from eDNA may be possible by detecting intraspecific genetic variation in mixtures of DNA and demonstrate the performance of this approach under multiple scenarios. We conclude by discussing key knowledge gaps that need be addressed to harness the full potential of eDNA-based

intraspecific genetic variation to inform questions in ecology, evolutionary biology, and conservation management.

Introduction

The use of molecular surveillance approaches such as environmental DNA (eDNA) to detect species has become more prevalent in recent years due to improved sensitivity and efficiency compared with conventional sampling methods (Fediajevaite, Priestley, Arnold, & Savolainen, 2021). Through the analysis of the genetic material present in environmental samples, eDNA approaches allow for the detection and quantification of single species (Ficetola, Miaud, Pompanon, & Taberlet, 2008; Jerde, Mahon, Chadderton, & Lodge, 2011; Ruppert, Kline, & Rahman, 2019) or entire species assemblages (Deiner, Bik, et al., 2017, Valentini et al., 2016). The latter approach, called eDNA metabarcoding, involves the use of universal PCR primers aimed at amplifying a particular genetic region of multiple species simultaneously, followed by high-throughput sequencing and taxonomic assignment of DNA sequences to species. Because eDNA metabarcoding allows for cost-efficient, rapid, and noninvasive assessments of species richness, it is increasingly recognized as a useful alternative or complement to standard sampling approaches for use in short- and long-term biomonitoring programs, biodiversity assessments, conservation.

While eDNA metabarcoding approaches are well established for detecting organisms at or above the taxonomic level of species, genetic variation within the level of species is often disregarded. However, in recent years studies have emerged exploring the detection and quantification of genetic variation within species (i.e., intraspecific genetic variation) using eDNA approaches (see reviews in Adams et al., 2019; Sigsgaard et al., 2020). For instance, Uchii, Doi, and Minamoto (2016) identified a single nucleotide polymorphism (SNP) in the mitochondrial genome that distinguishes invasive populations of Asian Carp from native populations and measured the relative abundance of each genetic variant in eDNA samples

throughout Japan. Sigsgaard et al. (2016) sequenced Whale Shark mitochondrial DNA haplotypes from eDNA samples and established that the read frequencies of haplotypes in eDNA samples align with Whale Shark haplotype frequencies from public databases and sequenced tissues. These early studies demonstrate that it is now possible to detect intraspecific genetic variation from eDNA, opening up new lines of opportunity to advance genetics tools to inform ecology, evolution, and conservation.

To date, analysis of intraspecific genetic variation from eDNA samples has been restricted to the detection and analysis of haplotypes in the short regions of the mitochondrial genome that are typically targeted in eDNA metabarcoding (Sigsgaard et al., 2020). However, by targeting other genetic marker types, including longer and/or more variable mitochondrial and nuclear markers, eDNA approaches have the potential to uncover much more detailed population genetic patterns and may produce improved estimates of species abundance. Regardless of the target genetic region, analysis of population-level genetic variation from eDNA samples has several key limitations that will require careful consideration to avoid drawing incorrect inferences. In this paper, we discuss the challenges and opportunities to study intraspecific genetic variation using eDNA. Specifically, we discuss the differences among genetic marker types and how they differ with regard to detectability, information content, and types of analyses that may be possible with eDNA-derived data. We also review the major causes of errors in eDNA data, discuss their consequences for population-level analyses, and outline bioinformatic approaches to detect and remove erroneous sequences. Finally, we present an approach for using intraspecific genetic variation to estimate species-specific abundance from eDNA and test the performance of this model using different genetic markers. By highlighting recent developments and future opportunities to study intraspecific

genetic variation with eDNA, we provide a framework for researchers interested in using eDNA to gain insights into the population characteristics of species.

Choice of molecular genetic marker

As in conventional specimen tissue-based population genetics research, the choice of molecular genetic marker targeted in eDNA research should be made by considering both practical criteria and the biological questions at hand (Anne, 2006). To deduce patterns of intraspecific genetic variation from eDNA, the targeted genetic region must contain sufficient levels of genetic variation at frequencies that can be detected using eDNA approaches. Accordingly, key considerations in the selection of genetic targets for eDNA research include the target amplicon size as well as the type and number of molecular genetic markers.

Marker length

While longer reads are more likely to contain sequence variation, longer DNA fragments degrade more rapidly and are less abundant in environmental samples (Bylemans, Furlan, Gleeson, Hardy, & Duncan, 2018; Wei, Nakajima, & Tobino, 2018). Because of this, many eDNA metabarcoding studies rely on the detection of shorter DNA fragments (< 200 bp) to characterize species (Bylemans, Gleeson, Hardy, & Furlan, 2018; Miya et al., 2015; Valentini et al., 2016). However, several studies have shown that longer amplicons can be reliably detected using eDNA approaches (Deiner, Walser, Mächler, & Altermatt, 2015; Evans et al., 2016; Jo et al., 2017), and due to their enhanced ability to discern closely related species, longer barcodes detect more species in eDNA samples than shorter barcodes (Zhang, Zhao, & Yao, 2020). When analyzing intraspecific genetic variation, longer DNA fragments

are similarly preferable to shorter fragments, as more genetic variation is likely to be detected within longer amplicons, allowing for higher resolution population analyses to be performed.

Although short-read high-throughput sequencers such as Illumina platforms are commonly used in eDNA metabarcoding research due to their high accuracy and low cost, they can only generate contiguous reads up to 600 base pairs (Heather & Chain, 2016). Long-read sequencing technologies (i.e., Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing) support much longer read lengths, routinely exceeding 10 Kb (Goodwin, McPherson, & McCombie, 2016). However, a key limitation of long-read sequencing is that it provides lower base-calling accuracy than short-read sequencing (Jain et al., 2015), and highly accurate reads are required when investigating intraspecific genetic variation from eDNA. Although error rates have declined in recent years (Amarasinghe et al., 2020), consensus sequencing strategies and bioinformatic tools are still necessary to detect and remove DNA sequencing errors (Baloğlu et al., 2021; Calus, Ijaz, & Pinto, 2018; Li et al., 2016). Further laboratory and field experiments are needed to determine the detectability of longer DNA fragments in environmental samples; however, if long DNA fragments can be reliably amplified and error rates can be adequately addressed, long-read sequencing may be a promising avenue for investigating intraspecific genetic diversity using eDNA approaches.

While most current approaches typically rely on PCR amplification of small sections of target genes, PCR-free approaches may also be possible, allowing for the sequencing of much longer genomic regions. For instance, nontargeted shotgun metagenomic sequencing of eDNA samples may allow for the quantification of multiple genetic regions or even full mitochondrial nuclear genomes (Farrell et al., 2022). However, the indiscriminate nature of

this approach makes it highly inefficient, with non-target taxa dominating shotgun sequence reads (Stat et al., 2017). Therefore, unless highly targeted sample collection methods are used (e.g., collecting samples in footprints Farrell et al., 2022; Székely et al., 2021), environmental shotgun sequencing is unlikely to provide enough reads from the target species to characterize population level diversity. Alternatively, eDNA samples may be enriched for specific genetic regions or full genomes of the target species using synthetic DNA probes designed from reference sequences (Dowle, Pochon, C. Banks, Shearer, & Wood, 2016; Wilcox et al., 2018). Although this approach may be challenging when starting concentrations of target species DNA are very low (Pinfield et al., 2019), recent work has demonstrated promise in target capture for sequencing full mitogenomes and hundreds of nuclear loci from aquatic eDNA samples (Jensen et al., 2021). Thus, while PCR-free methods for sequencing long reads and full genomes may be possible, the efficiency and reliability of such approaches for revealing intraspecific genetic variation from eDNA warrant further exploration.

Genetic marker type

In addition to marker length, genetic marker type is another key decision that will impact the ability to detect intraspecific genetic variants from eDNA. Nuclear and mitochondrial genetic material may vary widely in their respective concentrations within environmental samples, which in turn can impact genetic marker detectability during sample processing and analysis. This is because a typical eukaryotic cell contains hundreds to thousands of mitochondrial genomes versus a single nuclear genome (Cole, 2016), resulting in much higher expected mitochondrial DNA (mtDNA) concentrations in eDNA samples compared to nuclear DNA (nuDNA). To date, research on the ratio of mtDNA to nuDNA

concentrations in eDNA samples has focused on multi-copy nuclear ribosomal RNA (rDNA) genes such as 18S and ITS1, which are repeated nuclear genes that may have hundreds or thousands of copies per cell (Long & Dawid, 1980). Comparable copy numbers between mtDNA and rDNA markers is frequently reported, with rDNA genes sometimes exhibiting even higher detectability than mtDNA genes in eDNA samples (Dysthe, Franklin, McKelvey, Young, & Schwartz, 2018; Gantz, Renshaw, Erickson, Lodge, & Egan, 2018; Minamoto et al., 2017; Moushomi, Wilgar, Carvalho, Creer, & Seymour, 2019; Piggott, 2016). However, the ratio of mitochondrial to nuclear eDNA concentrations in environmental samples may change depending on a variety of factors including organism age, size, activity, tissue type, and environmental conditions that influence the production and degradation rates of different marker types (Bylemans, Furlan, et al., 2018; Bylemans et al., 2017; Furtwängler et al., 2018; Jo, Arimoto, Murakami, Masuda, & Minamoto, 2019, 2020). While most research on eDNA production, transport, and degradation is focused on mtDNA markers, the dynamics of eDNA can vary depending on the target gene, and the impact of such processes on eDNA detection and quantification for different marker types is not well understood.

Very few studies have documented the ability to amplify single-copy nuDNA markers (e.g., microsatellites and SNPs) from eDNA samples (but see Andres, Sethi, Lodge, & Andrés, 2021; Olson, Briggler, & Williams, 2012), and no direct comparison of marker detectability between single-copy nuDNA and mtDNA is yet available in the literature (but see Chapter 4 of this dissertation). However, several studies on the particle size distribution of macrobial eDNA have reported that captured DNA particles are frequently 1–10 μm in diameter, consistent with the hypothesis that eDNA is often intracellular, contained within cells or organelles (Jo et al., 2019; Turner et al., 2014; Wilcox, McKelvey, Young, Lowe, &

Schwartz, 2015). Therefore, although the copy number of nuDNA genes is expected to be much lower than mtDNA and rDNA, the collection of intact cells in environmental samples should make it possible to detect nuDNA microsatellites and SNPs. Improvements to eDNA collection, laboratory, and sequencing methodologies to increase nuDNA detectability may allow for the reliable amplification of nuDNA markers from environmental samples.

In addition to their relative concentrations in environmental samples, mtDNA and nuDNA markers differ in their properties (e.g., mutation rate, mode of inheritance, and degree of recombination), influencing the type and amount of genetic information they contain. Mitochondrial markers are the most common molecular marker used in eDNA research for research documenting diversity at the species-level and are to date the primary target genes in studies using eDNA to describe intraspecific diversity (Parsons, Everett, Dahlheim, & Park, 2018; Sigsgaard et al., 2020; Sigsgaard et al., 2016; Weitemier et al., 2021). However, while mtDNA barcoding regions may have high resolution at the species level, mitochondrial genes are generally maternally inherited, do not recombine, and are physically linked together. Single short mtDNA markers may not contain enough intraspecific genetic variation to conduct detailed population genetic analyses, and this limited information content cannot be overcome by targeting multiple mtDNA gene regions because they do not represent independent loci. If long mtDNA fragments or full mitochondrial genomes can be sequenced from eDNA samples (Deiner, Renshaw, et al., 2017), analyses of genetic structure and demographic history of a species may be possible. However, caution is warranted when drawing population inferences from mtDNA alone (Ballard & Whitlock, 2004), and many questions regarding intraspecific genetic variation may require the addition of nuclear genetic markers.

Ultimately, the molecular genetic marker selected to detect intraspecific genetic variation from eDNA will depend on the goals of the study, including the number of targeted taxa and the amount of genetic variation required to address the research questions (Figure 2.1). In the sections below, we review the types of mtDNA and nuDNA genetic markers that may be used to detect intraspecific genetic variation in eDNA and discuss the approaches that may be used to analyze these data.

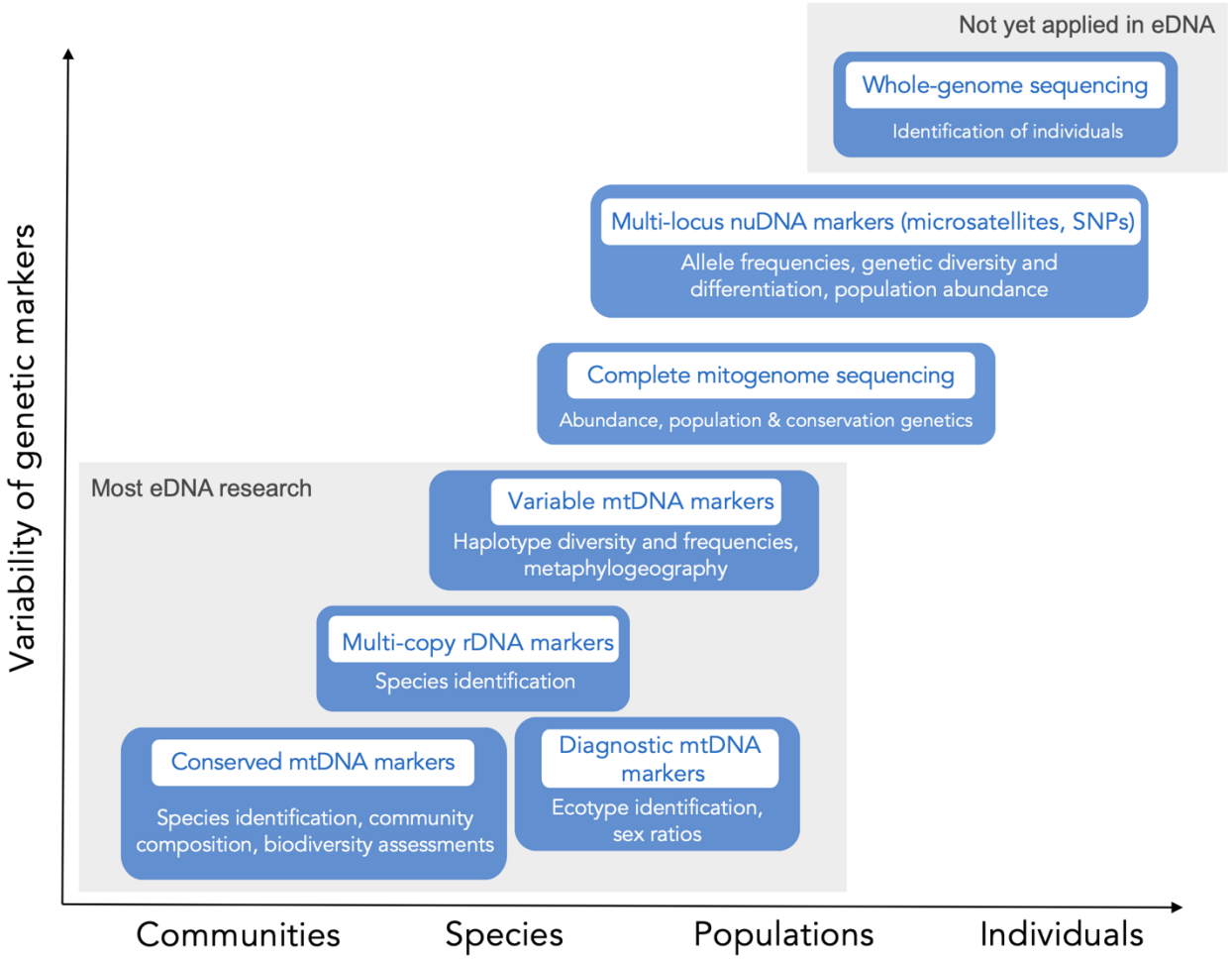

Figure 2.1. Molecular genetic markers used in eDNA research situated along axes of taxonomic resolution and the genetic variation contained in the targeted genetic markers. The

vast majority of eDNA research to date is aimed at characterizing communities or species using single or multi-locus mitochondrial markers. While a few studies have investigated within-species genetic variation using mitochondrial markers, more detailed patterns of population genetic diversity may be detected using longer and more variable gene regions or with nuclear eDNA.

Analyzing mitochondrial eDNA

In general, mitochondrial genes have a higher mutation rate and smaller effective population size than nuclear genes, allowing for the investigation of more recent demographic patterns. However, even with the high evolutionary rate of mtDNA markers, the short regions typically targeted in eDNA research may not contain enough genetic variation to be informative for population genetic analyses. Even so, the detection of intraspecific haplotypes and analysis of intraspecific genetic variation of different mtDNA markers represents an important step in eDNA-based population genetic analyses.

Barcoding genes (conserved mtDNA markers)

Mitochondrial genes are the marker of choice for DNA-based taxonomic identification, with the COI gene representing the most popular marker in DNA barcoding as the Barcode of Life (Hebert, Cywinska, Ball, & DeWaard, 2003; Ratnasingham & Hebert, 2007). Other gene regions commonly targeted for DNA barcoding of macro-organisms include 12 rRNA, 16 rRNA, ND2, and cytochrome B (cytb). DNA barcodes are specifically designed to maximize interspecific diversity while minimizing intraspecific diversity, and in most cases intraspecific sequence divergence in these gene regions is not expected to reach

levels required for detailed population genetic analyses. However, diagnostic sequence variants may be identified within mtDNA barcoding regions that can allow for the detection of specific haplotypes of interest. For instance, COI markers have been used for diagnostic identification of specific zebra and quagga mussel haplotypes (Marshall & Stepien, 2019), and a diagnostic cytb marker has been developed to distinguish among several closely related invasive carp species as well as among several silver carp haplotypes (Stepien, Snyder, & Elz, 2019). Diagnostic SNPs in the cytb gene have also been used to differentiate between black and white morphs of the salamander *Proteus anguinus* (Gorički et al., 2017). Therefore, while mtDNA barcoding genes may not contain high levels of genetic variation, they may be useful for the identification of *a priori* established genetic variants.

Variable mtDNA markers

The mitochondrial control region (i.e., d-loop) exhibits a higher mutation rate than other mtDNA markers, with a higher ratio of haplotypes to individuals making it a better candidate for intraspecific diversity analyses. Similar to barcoding genes, diagnostic SNPs within the d-loop have been used to differentiate between specimens from native and non-native carp populations (Uchii et al., 2016; Uchii, Doi, Yamanaka, & Minamoto, 2017) and to identify Killer Whale ecotypes (Baker, Steel, Nieukirk, & Klinck, 2018). However, more detailed population genetic information may also be uncovered. For instance, the relative read abundance of d-loop haplotypes can be estimated from eDNA samples in proportions similar to frequencies in the focal population (Parsons et al., 2018; Sigsgaard et al., 2016; Tsuji, Shibata, Sawada, & Ushio, 2020).

Full mitogenome

In addition to short mtDNA fragments, it may be possible to recover sequence data from entire mitochondrial genomes from eDNA samples (Deiner, Renshaw, et al., 2017; Jensen et al., 2021). However, although estimating full-genome mitochondrial variation from eDNA is possible, no study has yet explored the potential to estimate population genetic parameters from the resulting sequence data. Nonetheless, this area of research remains promising, as the sequencing of full mitogenomes indicates the persistence of intact organelles or cells, illuminating the possibility that entire cells could be isolated from environmental samples and sequenced individually.

Types of analyses with mtDNA data

An important consideration for all analyses of genetic variation from eDNA is that genetic variation is aggregated at the population level, representing a mixture of haplotypes from an unknown number of genetic contributors (but see Dugal et al., 2021). However, most existing population genetic equations, statistical software, and analytical frameworks are based on the availability of individual-level genotypes. Therefore, even if existing programs to estimate population genetic parameters can handle the intraspecific genetic datasets produced through the analysis of eDNA (e.g., haplotype frequencies), the assumptions and limitations of each analysis should be thoroughly investigated to ensure the robustness of the resulting population genetic inferences.

The ability to recover several haplotypes from eDNA allows for analyses to be conducted on the presence of different haplotypes or haplotype relative abundance (e.g., frequency). With data on the detection of haplotypes, parameters such as haplotype diversity,

nucleotide diversity, and segregating sites may be estimated for each sampled population. Rarefaction curves for the detection of haplotypes as a function of sample size may also be useful for predicting the total amount of genetic diversity at a site as well as the sampling effort that would be required to completely sample the diversity (Székely et al., 2021).

Haplotype frequencies from eDNA data may be inferred directly from eDNA sequence read frequencies, under the assumption that sequence abundance accurately reflects the abundance of the haplotype in the population. However, due to stochastic sampling, variable contributions of DNA from different individuals, and biases introduced during PCR or sequencing, this assumption may not hold true in many cases. More conservative alternatives to directly quantifying haplotype frequencies include to characterizing ranks of haplotype frequencies (Turon, Antich, Palacín, Præbel, & Wangensteen, 2020) or the presence/absence of haplotypes across replicated samples (Azarian, Foster, Devloo-Delva, & Feutry, 2021). Regardless of how haplotype frequencies are estimated, the data may be used to construct haplotype networks to illustrate the haplotype differences among sites or populations. Analysis of molecular variance (AMOVA) may also be conducted to determine the amount of genetic variation associated with each level of hierarchical organization of sampled sites. Other population genetic parameters, such as Fst and Tajima's D, may be calculated from population-level haplotype frequencies (Shum & Palumbi, 2021; Weitemier et al., 2021). However, such metrics are sensitive to the presence and frequency of haplotypes, and an investigation of how population parameters are impacted by the specific approach used to characterize haplotype frequencies from eDNA samples is needed.

An exciting avenue of research capitalizes on the understanding that eDNA samples contain a mixture of DNA from many species, making possible the evaluation of intraspecific

diversity from multiple species simultaneously (Shum & Palumbi, 2021; Stat et al., 2017; Turon et al., 2020; Weitemier et al., 2021). Although identifying haplotypes for entire communities may present additional challenges over single-species approaches, it may soon be possible to monitor genetic diversity for hundreds of species across the Tree of Life using eDNA approaches.

Analyzing nuclear eDNA

Nuclear genetic markers provide novel opportunities to explore a wider range of ecological and evolutionary hypotheses with eDNA, but few eDNA studies have employed the amplification nuclear markers. However, because multiple unlinked loci have higher information content and therefore greater power to detect genetic patterns, nuDNA markers may be a better choice than mtDNA for some population-level genetic analyses from eDNA.

Multi-copy rDNA markers

Nuclear ribosomal RNA (rDNA) genes such as 18S and ITS1 are present in multiple copies throughout the nuclear genome and therefore have higher detectability in eDNA samples than single-copy nuDNA and, in some cases, mtDNA (Minamoto et al., 2017). Although nuclear ribosomal genes may be useful as a barcode gene for differentiating closely related species (Toju, Tanabe, Yamamoto, & Sato, 2012), these genetic regions may only provide limited insight into intraspecific genetic variation (Booton et al., 1999). However, as with mtDNA barcoding genes, diagnostic sequence variants may be identified in these genes, making them useful for the detection and quantification of *a priori* sequence variants in eDNA.

Microsatellites

To date, Andres et al. (2021) is the only study to report the detection and quantification of multi-allelic microsatellite alleles from eDNA samples. In both laboratory and field experiments, allele frequencies estimated from sequencing eDNA samples with a panel of 28 Round Goby (*Neogobius melanostomus*) microsatellite loci closely resembled allele frequencies from genotyped issues, suggesting that population genetic parameters may be estimated from sequence read counts from environmental samples. As with detecting mtDNA haplotype variation, proper identification of sequence length variants (i.e., microsatellite alleles) from eDNA samples will likely require next-generation sequencing approaches. Although microsatellite alleles from eDNA samples could potentially be identified using peak heights in eletropherograms, the complex nature of mixed DNA samples may make it challenging to decipher alleles (Sethi, Larson, Turnquist, Isermann, & Kembel, 2019). Allele identification based on size alone may also mask genetic variants due to substitutions (i.e., SNPs) that may be present in the same amplified region. For these reasons, identifying alleles using sequencing-based approaches are the preferred method for quantifying microsatellite alleles from eDNA samples.

As described in the mtDNA section above, the estimation of allele frequencies from eDNA may be conducted using quantitative (read frequencies) or semi-quantitative (read frequency rank; presence/absence of alleles in replicates) approaches. Once allele frequencies are estimated, several population genetic characteristics may be quantified. For instance, allelic richness, expected heterozygosity, and the number of private alleles can be estimated and compared among populations, and population genetic structure may be determined using

AMOVA or principal components analysis (PCA) of allele frequencies. Genetic distances among populations may also be calculated using the allele frequency distance (AFD), a metric that is highly correlated with common genetic differentiation metrics such as Fst but does not require individual genotypes (Berner, 2019).

SNPs

While the per-locus information content of bi-allelic SNPs is lower than that of multi-allelic microsatellites, SNPs are abundant throughout the nuclear genome, and dozens or hundreds of SNP loci may be targeted at once. To facilitate the amplification of high numbers of loci, SNP assays may be multiplexed, where several SNP loci are co-amplified in a single PCR. However, primers in multiplex PCR may interact and lead to inhibition of some loci, a problem that may worsen when starting concentrations of template DNA are variable as is expected to be the case in eDNA samples. Careful optimization of multiplex PCR is therefore recommended (Elnifro, Ashshi, Cooper, & Klapper, 2000).

If several independent SNPs can be amplified from eDNA samples, the analysis of SNPs may be possible within pooled sequencing (Pool-seq) approaches (Gautier et al., 2013; Hivert, Leblois, Petit, Gautier, & Vitalis, 2018; Kofler, Pandey, & Schlotterer, 2011). These approaches have been developed for the sequencing of pooled DNA from many individuals, a cost- and time- effective alternative to individual genotyping. Analytical frameworks have been developed to estimate allele frequencies and population genetic parameters such as Fst from Pool-seq data, and several pipelines for the analysis of pooled sequencing data are already available (Kofler et al., 2011). Most Pool-seq analytical frameworks are developed

only for bi-allelic loci, and while useful for analysis of SNPs, these frameworks are not applicable to multi-allelic markers such as microsatellites.

Although a single SNP may not contain high information content, several SNPs may be present within a short gene region that can be amplified and sequenced to general multi-allelic markers. These "microhaplotype" markers contain more information content per locus, allowing for more powerful genetic analyses to be conducted using fewer loci (Baetscher, Clemento, Ng, Anderson, & Garza, 2018; Kidd et al., 2014). However, such multi-allelic markers may no longer be applicable in most Pool-seq analysis pipelines, many of which assume independence among loci, an assumption that is violated when SNPs are located in close proximity (Slatkin, 2008). In this sense, analysis of microhaplotype loci may be more similar to that of microsatellite loci.

While analysis of highly variable mitochondrial and nuclear genetic markers is not yet common in eDNA research, we believe eDNA approaches tailored for this purpose offer an opportunity to study detailed population genetic patterns, potentially with higher sensitivity at a lower cost. However, validation experiments and special considerations will be necessary to fully understand the potential and limitations of population genetic analysis of eDNA samples. In the section below, we summarize the importance of considering false positive and false negative detections and the resulting impact on the precision of estimates of population genetic parameters.

Understanding and accounting for errors

Regardless of the length or type of marker selected, a requisite step in any eDNA study is the detection of spurious DNA sequences in order to distinguish authentic DNA

sequences from errors. This step becomes even more important in the analysis of intraspecific genetic variation from eDNA, as the detection and quantification of exact sequence variants is the primary goal. At the same time, eDNA samples often contain DNA in low quantities and quality, and the resulting sequences may be more prone to errors than those from high-quality DNA. Erroneous genetic variants may have substantial implications for downstream population genetics analyses. Therefore, understanding the consequences of sequencing errors and testing approaches to remove errors while retaining true low-frequency genetic variants presents a priority area for statistical and bioinformatics applications in eDNA research.

Causes and consequences of errors in eDNA sequence data

Addressing erroneous sequence variants in eDNA data first requires an understanding of how such errors arise. Furlan, Davis, and Duncan (2020) provide a comprehensive overview of errors in metabarcoding workflows and classify them as either contamination errors or misidentification errors that can be generated during collection, amplification, sequencing, or bioinformatics processes. Because errors in eDNA data can arise from multiple complex and potentially interacting origins, the nature and severity of errors may differ depending on the specific study conditions. Interpreting the detection of intraspecific genetic variation in eDNA metabarcoding data requires particular consideration of the causes and consequences of such errors in different contexts.

Contamination errors and misidentification errors (i.e., erroneous DNA sequences or incorrect sequence assignment) generally lead to false positive detections and an overestimation of intraspecific genetic diversity in a sample, with implications for downstream analyses of genetic diversity, differentiation, and species abundance estimates.

On the other hand, rare genetic variants that are present at very low frequencies in eDNA could be missed during field sampling or during laboratory procedures, leading to leading to false negatives and an underestimation of genetic diversity. Stringent protocols to filter out erroneous sequences can also result in the removal of true DNA fragments, particularly if the abundance of rare genetic variants does not exceed the abundance of erroneous sequences. While the sensitivity and efficiency of different assays should be investigated to determine bioinformatic thresholds, a trade-off between decreasing the rate of false positives and increasing the rate of false negatives may persist. Several approaches to detect and remove PCR and sequencing errors have been developed (Schnell, Bohmann, & Gilbert, 2015), but a better understanding of these specific processes may be needed to increase confidence in delineating genetic variation within species from eDNA.

Identification and removal of erroneous sequences

Removing errors in eDNA data may be accomplished through the use of bioinformatic thresholds and algorithms designed to identify and eliminate sequences that are likely to be erroneous. The selection of specific bioinformatic parameters may vary by study and should be made with the study goals, data type, and implications of errors in mind.

Assuming that erroneous sequences will be present at low frequencies compared to authentic sequences in eDNA samples, setting minimum abundance thresholds to remove low-frequency sequence variants may be a straightforward approach to filtering out errors. Several studies use the estimated error rate of a particular gene region or sequencing platform as a threshold that any given sequence variant must exceed to be retained in the dataset (Sigsgaard et al., 2016; Stat et al., 2017). If using multiple loci with variable numbers of

alleles per locus, a variable threshold based on per-locus allelic richness may be required to obtain more accurate allele frequency estimates, as erroneous sequences may consume a greater percentage of reads when the genetic complexity of a sample is low (Andres et al., 2021). Other threshold-based approaches may include retaining only the alleles that are observed in multiple field or laboratory replicates, under the assumption that the probability of random errors is low and observing any allele in multiple replicates becomes exceedingly rare. However, for the same reason, the probability of detecting of low-frequency genetic variants in multiple replicates is low and such thresholds may exclude true genetic variants.

To reduce the possibility of amplification of non-target species, applying a sequence similarity threshold or alignment algorithm to known (query) sequences may be desirable (e.g., Smith-Waterman; BLAST; Altschul, Gish, Miller, Myers, & Lipman, 1990; Smith & Waterman, 1981). However, to maximize confidence that the variation detected in eDNA is biologically real, it may be beneficial to restrict the list of alleles to only those known from high quality tissue-based genotyped samples (as in Parsons et al., 2018; Shum & Palumbi, 2021). Such a conservative approach will remove nearly all errors due to PCR/sequencing error or non-target amplification but may result in an underestimate of genetic diversity, as any genetic variants that have not been previously identified will not be retained.

While threshold-based exclusion of sequences can effectively remove low-frequency errors, several pipelines have been developed to correct sequencing errors and determine real biological sequences, also referred to as amplicon sequence variants (ASVs). Such denoising approaches, implemented in algorithms such as DADA2 (Callahan et al., 2016), UNOISE2/3 (Edgar, 2016), and Deblur (Amir et al., 2017), allow for more refined detection and removal of spurious sequences. Analyses of bioinformatic pipelines have shown DADA2 to be highly

effective in removing sequencing errors while retaining biological sequence variants, particularly when steps to remove sequence chimeras are added (Macé et al., 2022; Tsuji, Maruyama, et al., 2020; Tsuji, Miya, et al., 2020). Other approaches to determine erroneous sequences include an examination of entropy changes in the different codon positions, an approach that may be used if variation in coding sequences is being assessed (Turon et al., 2020). For nuclear genetic markers to become more useful in eDNA research, future pipelines will need to identify and remove errors specific to nuDNA (e.g., microsatellite stutter).

Bioinformatic pipelines to correctly identify and remove erroneous DNA sequences may combine error correction, sequence similarity clustering, and threshold-based approaches depending on the complexity of the dataset and study goals. To demonstrate the behavior of different error removal approaches, we examined the number of genetic variants remaining in an eDNA dataset after applying a sequence of increasingly stringent bioinformatic filters. We reanalyzed data from Andres et al. (2021) that contains Illumina sequence data for 28 multi-allelic Round Goby microsatellite loci from eDNA samples collected in Cayuga Lake, NY (Figure 2.2). Raw sequence data contained an average of 335 sequence variants per locus, but only 53 of these variants exceeded 1% of the sequence reads in a sample (Figure 2.2). The DADA2 algorithm (including a chimera removal step) removed more sequences than the abundance-based threshold, but examination of the ASV sequences revealed possible amplification of non-target species at several loci. Therefore, we applied the Smith-Waterman clustering algorithm following DADA2 to remove sequences outside of a sequence similarity threshold to a parent sequence based on genotyped tissues from 73 specimens. Simply restricting ASVs detected in eDNA to alleles known from genotyped tissues resulted in the most extreme exclusion of genetic variants, with only 3.1 alleles per locus remaining (Figure

2.2). While it is impossible to know which variant number is most accurate, we suspect that it is somewhere between the lowest two estimates, as some rare alleles may not have been present in the 73 genotyped specimens may have been detected in eDNA samples.

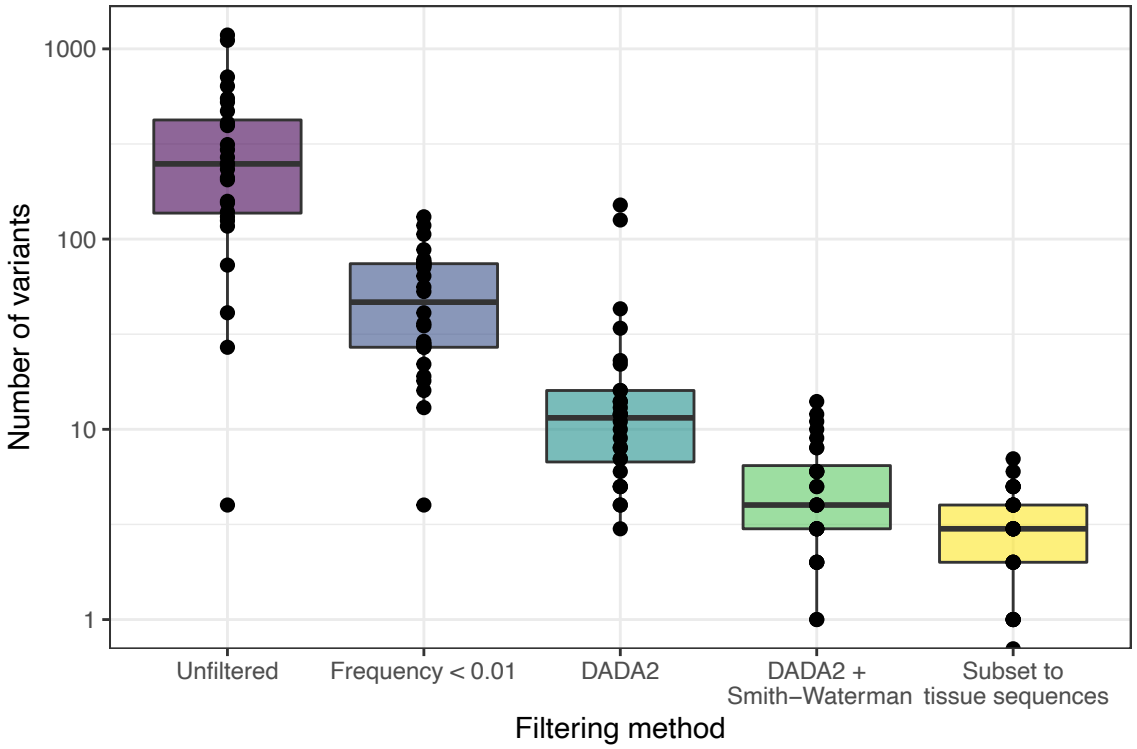

Figure 2.2. Boxplot distribution of the number of amplicon sequence variants (ASVs) generated per locus with 1) unfiltered sequencing output; 2) removal of all ASVs with frequency <1% in a sample; 3) DADA2 denoising algorithm; 4) DADA2 denoising algorithm with Smith-Waterman sequence alignment to a tissue sequence; and 5) sub-setting eDNA sequences only to variants found in specimen genotypes. The number of variants remaining for each of the 28 microsatellite loci are shown in black circles.

Preventing errors and minimizing their effects in eDNA analyses

Some of the challenges with distinguishing true genetic variation from background noise may be addressed by minimizing erroneous sequences and increasing the quality and quantity of target DNA captured in environmental samples. Several best practices to prevent and account for contamination errors in the field and laboratory are now commonplace in eDNA research, including decontaminating sampling equipment, processing samples in a dedicated clean laboratory, adhering to sterile laboratory practices, and collecting negative controls (i.e., blanks) at each stage of the process (Goldberg et al., 2016).

Novel filtration approaches allowing for the collection of large sample volumes (e.g., up to 3,000 L of water) have exhibited improved detection of rare species (Sepulveda et al., 2019), and similar methods will likely improve the detection of rare genetic variants from eDNA. Sampling at the locations or times of the year where focal species are known to occur may also be employed to increase the concentration of target species' DNA in a sample, as in the case of sampling in the wake of marine mammals (Baker et al., 2018; Parsons et al., 2018; Sigsgaard et al., 2016) or during known feeding or spawning aggregations. Improvements to eDNA extraction, amplification, and sequencing efficiency may also be able to overcome some of the technical limitations of detecting intraspecific genetic variation in eDNA, and the potential for such technological improvements should be explored.

Although increasing the number of PCR cycles may compensate for lower starting concentrations in eDNA samples, high numbers of PCR cycles may also reduce the detectability of taxa depending on amplification efficiency and their relative rarity in a sample

(Kelly, Shelton, & Gallego, 2019). We expect the number PCR cycles and amplification efficiencies will similarly impact intraspecific diversity metrics derived from eDNA samples.

Even if all measures are taken to prevent, identify, and remove spurious DNA sequences, it may remain difficult to distinguish between erroneous sequences and true low-frequency alleles in the population using eDNA approaches. In such cases, error models to account for the introduction of false alleles as well as the non-detection of rare alleles may be able to increase the reliability of eDNA approaches by estimating the frequency of sequence artifacts and taking into account stochastic sampling of DNA molecules (Miller, Joyce, & Waits, 2002; Taberlet et al., 1996). The impact of false positives, false negatives, and allele frequency misspecifications on the estimation of population genetic parameters from eDNA data must also be investigated. For instance, several studies have examined the consequences of sequencing and genotyping errors on biological inferences in eDNA research and non-invasive genotyping approaches, revealing the bias in genetic parameters induced under different numbers and types of errors (Burian et al., 2021; Hivert et al., 2018; Pompanon, Bonin, Bellemain, & Taberlet, 2005). Similar approaches will be useful for understanding how different errors may lead to biased estimates of population genetic parameters from eDNA.

Estimating abundance using eDNA

eDNA copy number as a metric of abundance

In addition to the molecular detection of organisms, eDNA approaches may be useful for estimating organismal abundance, a metric that is central to species monitoring, management, and conservation (Yoccoz, Nichols, & Boulinier, 2001). To date, most efforts to

assess species abundance with eDNA have focused on correlating ambient species-specific eDNA concentration with numerical abundance, with several studies reporting a positive correlation between the two metrics (Baldigo, Sporn, George, & Ball, 2017; Doi et al., 2017; Lacoursière-Roussel, Côté, Leclerc, & Bernatchez, 2016; Pilliod, Goldberg, Arkle, & Waits, 2013; Schmelzle & Kinziger, 2016). The empirical relationship between eDNA concentration and species abundance reflects an assumption that all individuals from a species present at a sampling location produce similar amounts of DNA and that eDNA degrades consistently with respect to environmental conditions and DNA source. However, in many instances this assumption is unlikely to be true, and such relationships may be confounded by variation in eDNA concentration due to biotic and abiotic factors influencing DNA production and loss rates as well as biases introduced through the use of different eDNA sampling, processing, and amplification approaches (Beng & Corlett, 2020).

The amount of eDNA captured in a sample is the collective result of several complex processes influencing eDNA shedding, transport, and decay, settling, and resuspension, all of which may occur at different rates under different environmental conditions including temperature, pH, UV exposure, microbial activity, flow regime, and substrate type (Barnes & Turner, 2015; Carraro, Hartikainen, Jokela, Bertuzzo, & Rinaldo, 2018; Shogren et al., 2017; Strickler, Fremier, & Goldberg, 2015). The production rate of eDNA can also vary among source individuals as a function of their size, behavior, or metabolism, further obscuring the relationship between eDNA concentration and organism abundance (Dunn, Priestley, Herraiz, Arnold, & Savolainen, 2017; Klymus, Richter, Chapman, & Paukert, 2015; Maruyama, Nakamura, Yamanaka, Kondoh, & Minamoto, 2014). Without accounting for these sources of

variation, eDNA concentration may not accurately reflect the abundance of the focal species, particularly at local scales.

The type and size of molecular marker targeted may also impact variation in eDNA copy numbers and distort the link between DNA concentration and species abundance. For instance, while the number of nuDNA sequences per cell do not vary among source tissue types, mtDNA copy number per cell can be highly variable within and among individuals (Long & Dawid, 1980; Robin & Wong, 1988). As discussed above, the decay rate of eDNA depends on the target fragment length and source (i.e., mtDNA or nuDNA), which can impact the detection and yield of target species DNA (Jo et al., 2020). eDNA concentration may also be sensitive to eDNA sampling and processing approaches including volume of water filtered, filter type, preservation buffer, filter membrane type, extraction method, quantification approach (i.e., quantitative PCR, digital droplet PCR), and primer specificity (Goldberg et al., 2016). Although there have been attempts to improve the methodological and reporting standards in eDNA research, the lack of standardized protocols and analytical frameworks for evaluating the quantity of eDNA prevents comparisons of eDNA-based species abundance estimates among studies and systems (Loeza-Quintana, Abbott, Heath, Bernatchez, & Hanner, 2020).

Although several studies have improved our understanding of eDNA dynamics under different environmental conditions and protocols, it is often not feasible to account for all of the potential sources of variation in eDNA concentrations, particularly in complex natural environments. A recent meta-analysis examined the studies reporting correlations between the concentration of species-specific eDNA particles and the density or biomass of a species, and found that on average, eDNA particle concentration accounted for 57% of the observed

variation in species abundance in natural systems (Yates, Fraser, & Derry, 2019). A better understanding of the sources of variation in eDNA concentration will be a requisite step in obtaining robust estimates of species abundance based on eDNA concentrations. However, identifying and controlling for this variation in natural systems may not be possible, and the utility of correlation-based approaches may remain limited due to high levels uncertainty. We propose that some of this uncertainty may be addressed through the use of alternative analytical frameworks that incorporate information about genetic diversity in eDNA samples to estimate species abundance.

Genetic variation as a metric of abundance

Rather than focusing on the concentration of target species eDNA, the amount of genetic variation contained in an eDNA sample may be used to assess species abundance by estimating the absolute number of individuals contributing DNA to the mixture. In the simplest sense, the minimum number of individuals detected in an eDNA sample is equivalent to the number of individuals required to produce the observed set of genetic variants. That is, the number of genetic contributors to a DNA mixture must be at least the number of detected genetic variants at a locus divided by the ploidy of the genetic marker. For instance, the number of individuals in a haploid mtDNA mixture must be greater than or equal to the number of distinct haplotypes/1, whereas the number of individuals in a diploid nuDNA mixture must be greater than or equal to the number of detected alleles/2 (Carreon-Martinez, Wellband, Johnson, Ludsin, & Heath, 2014). This approach, termed the maximum allele count, is useful for providing a lower bound on the number of contributors required to explain the observed set of haplotypes or alleles. However, it does not account for instances where

multiple individuals may share the same genotype, and thus tends to underestimate the number of contributors in mixed DNA samples, a bias that worsens as the number of individuals increases (Haned, Pene, Lobry, Dufour, & Pontier, 2011; Paoletti, Doom, Krane, Raymer, & Krane, 2005). This problem is particularly severe when targeting genetic markers that exhibit a highly skewed frequency spectrum, as it is more likely that sampled individuals will exhibit the same common genotypes.

The issue of redundant alleles may be addressed through the use of likelihood-based models, where the observed alleles in a mixture and the allele frequencies of the focal population are used to calculate the likelihood that a putative number of individuals (x) produced an observed set of alleles (A) given the population allele frequencies (p) of the observed alleles (Egeland, Dalen, & Mostad, 2003; Haned et al., 2011; Sethi et al., 2019). This likelihood can be calculated and multiplied across any number of unlinked loci, with the estimated number of individuals determined by the maximum likelihood estimate across any given number of putative contributors. This approach originated in the forensic sciences and the applications of DNA mixture models were extended into ecological frameworks by Sethi et al. (2019), who demonstrated the performance of the model using simulated mixtures of different marker types as well as the ability to estimate yellow perch predation rates through stomach content analyses. Andres et al. (2021) subsequently demonstrated potential for estimating abundance from DNA mixtures using eDNA samples collected from experimental mesocosms.

Estimating abundance using genetic diversity eliminates the challenge of variable DNA production rates within and among individuals by estimating the number of distinct individuals contributing DNA to an environmental sample. Because it does not rely on DNA

copy number, this approach is not influenced by differences in eDNA production due to organism body size, behavior, metabolism, or genetic marker type. While Sethi et al. (2019) and Andres et al. (2021) successfully used DNA mixture models to estimate the number of genetic contributors when the total number of individuals is small (≤ 10), the limitations of this approach have not been thoroughly explored in highly complex mixtures of greater numbers of individuals. Below, we explore the ability of DNA mixture models to resolve mixtures of large numbers of individuals to better understand the applicability of this approach in natural systems, opening the door for avenues for future research and application.

Simulating DNA mixtures with real genotypes

To advance understanding of the performance of DNA mixture models in ecological and environmental DNA applications, we explored abundance estimation in simulated mixtures of DNA made up of larger numbers of putative contributors that may reflect local abundances experienced in ecological studies. We also explored mixture performance under a suite of genetic marker types including novel demonstration of mtDNA markers. We used genotypes from a published dataset that contains microsatellite and SNP marker data for 1,129 Linesnout Gobies (*Elacatinus lori*) collected from 35 sites across the Belize Barrier Reef (for detailed collection, laboratory, and sequencing methods, see D'Aloia et al., 2020). We also used mitogenome sequence data (~ 500 bp contigs) that were collected for each of these individuals (unpublished data; see Supplementary Methods S2.1). Our goal was to generate artificial DNA mixtures using combinations of real genotypes for different molecular marker types, numbers of loci, and numbers of individuals used to calculate population allele frequencies.

Due to the high reported levels of genetic structuring in this region, we selected only the 127 individuals from four sites at one of the subpopulations detected in D'Aloia et al. (2020). Furthermore, the DNA mixture abundance model is combinatorial in nature and estimating the number of contributors can become prohibitively computationally intensive with large numbers of alleles at loci. Thus, we excluded highly multi-allelic microsatellite loci (i.e., > 20 alleles) from the dataset to reduce computation times in analyzing simulated mixtures. We assembled eight mitochondrial sequence contigs to generate haplotype data for a 4 kB mtDNA fragment, where any sequence variations within the 4 kB region were designated as a unique mitochondrial haplotype. The final dataset included individual genotypes for 44 microsatellite loci, 256 nuclear SNP loci, and mitochondrial haplotype information for 127 individuals.

The DNA mixture model requires an observed list of alleles (A) from a mixture sample and their associated population allele frequencies (p) as inputs in estimating sample abundance. We generated DNA mixture genotypes by randomly sampling n individuals from each molecular marker dataset, with n ranging from 2–100 individuals (i.e. *in silico* mixtures made up of 2–100 contributors). Genotypes from sampled individuals were combined to generate a list of observed alleles (A) as a simulated DNA mixture. In practice, input allele frequencies for mixture estimation can be generated from a reference set of tissue-based samples from specimens in a population, although Andres et al. (2021) also showed that it may be possible to directly infer population allele frequencies from read counts from eDNA samples as well. Accuracy of the DNA mixture model may depend on the presence of rare alleles, whereby rare alleles are generally strongly informative in characterizing mixture abundance (Sethi et al., 2019). By their nature, however, characterizing the frequencies or rare

alleles from samples of specimens from a population may be more difficult. This is because the number of individuals used to calculate population allele frequencies is related to the rarity of the alleles that can be detected in a mixture. For instance, the rarest microsatellite allele frequency that could be detected in a dataset of 25 diploid individuals is 0.02 (1 allele copy in 50 total allele copies), while in a dataset of 100 diploids it is 0.005 (Table 2.1). We therefore explored the impact of assessing allele frequencies form relatively small (25) or large (100) reference set of individuals.

The number of loci used in the mixture may also influence model accuracy (Sethi et al., 2019). We estimated the number of individuals in each DNA mixture using different numbers of nuclear loci to represent a relatively small vs. large marker panel (10 vs. 44 loci for microsatellites; 64 vs. 256 loci for SNPs). Because mitochondrial DNA haplotypes were generated by concatenating multiple contigs, we ran the DNA mixture model using mitochondrial haplotypes from a single mitochondrial contig (480 bp) in addition to the full 4 kB fragment. In sum, the DNA mixture model was used to estimate the number of contributors in simulated mixtures ranging in size from 2–100 individuals for each of three marker types (microsatellites, SNPs, and mitochondrial), with population allele frequencies specified using different numbers of individuals (25 or 100), using either a small or large panel/fragment size for each marker type.

Table 2.1. Mean bias (± 1 s.d.) in the estimated number of contributors in simulated DNA mixtures of True N = 10, 40, and 100 individuals. Mixtures were constructed using *E. lori* genotypes for microsatellites, SNPs, and mitochondrial haplotypes. The number of individuals

in each mixture was estimated using relatively large or small numbers of loci and population allele frequencies were estimated using either 100 or 25 individuals.

Marker type	No. of individuals for allele freq.	No. of loci[1]	No. of alleles	Frequency of rarest allele	Mean bias *N*=10	Mean bias *N*=40	Mean bias *N*=100
Microsatellite	100	44	460	0.005	-0.1 ± 0.3	4.1 ± 2.8	1.4 ± 8.5
	25	44	402	0.020	-1.9 ± 0.3	-16.7 ± 1.1	-63.7 ± 4.9
	100	10	121	0.005	-0.6 ± 1.4	2.4 ± 5.6	-0.7 ± 17.1
	25	10	91	0.020	-2.4 ± 1.2	-16.0 ± 4.3	-73.2 ± 0.8
SNP	100	256	512	0.005	0.9 ± 1.7	5.4 ± 6.2	35.5 ± 4.9
	25	256	512	0.020	-1.6 ± 1.1	-7.3 ± 4.1	-21.8 ± 13.0
	100	64	128	0.005	0.6 ± 2.54	35.1 ± 35.5	12.7 ± 31.8
	25	64	128	0.020	2.7 ± 3.4	8.7 ± 11.1	-37.0 ± 0.0
Mitochondrial	100	8	40	0.010	-0.7 ± 1.7	0.4 ± 7.0	–
	25	8	20	0.040	-4.7 ± 1.3	-16.8 ± 4.6	-42.9 ± 10.4
	100	1	23	0.010	2.3 ± 2.5	1.9 ± 9.2	-70.3 ± 29.6
	25	1	8	0.040	-3.9 ± 2.1	-16.2 ± 6.8	-60.0 ± 0.0

[1] Numbers for mitochondrial data represent the numbers of contigs used to generate mitochondrial haplotypes rather than independent loci.

A challenge with mixtures made up of large numbers of contributors is that loci may 'saturate' insofar that all possible alleles manifest in a sample. In such cases when all possible alleles are observed at a locus, the likelihood that a putative number of contributors produces an observed set of alleles will monotonically increase, and, when all alleles are observed across all loci, a maximum likelihood estimate cannot be reached (Egeland et al., 2003). Furthermore, because the likelihood is taken as a product across all loci, the maximum likelihood may be upwardly biased when some loci exhibit a monotonically increasing likelihood (i.e., all alleles are observed), even if other loci exhibit a single maximum likelihood (Figure S2.1). To address these maximum likelihood estimation challenges for large mixtures, we removed any microsatellite loci for which all alleles were observed when estimating the number of contributors to a sample DNA mixture. While this filtering step

works well for multi-allelic loci such as microsatellites, filtering out saturated loci was not possible for SNPs, as both alleles for most loci are frequently observed even in small mixtures of DNA. This was also not possible for mitochondrial haplotype data, as only a single locus was used.

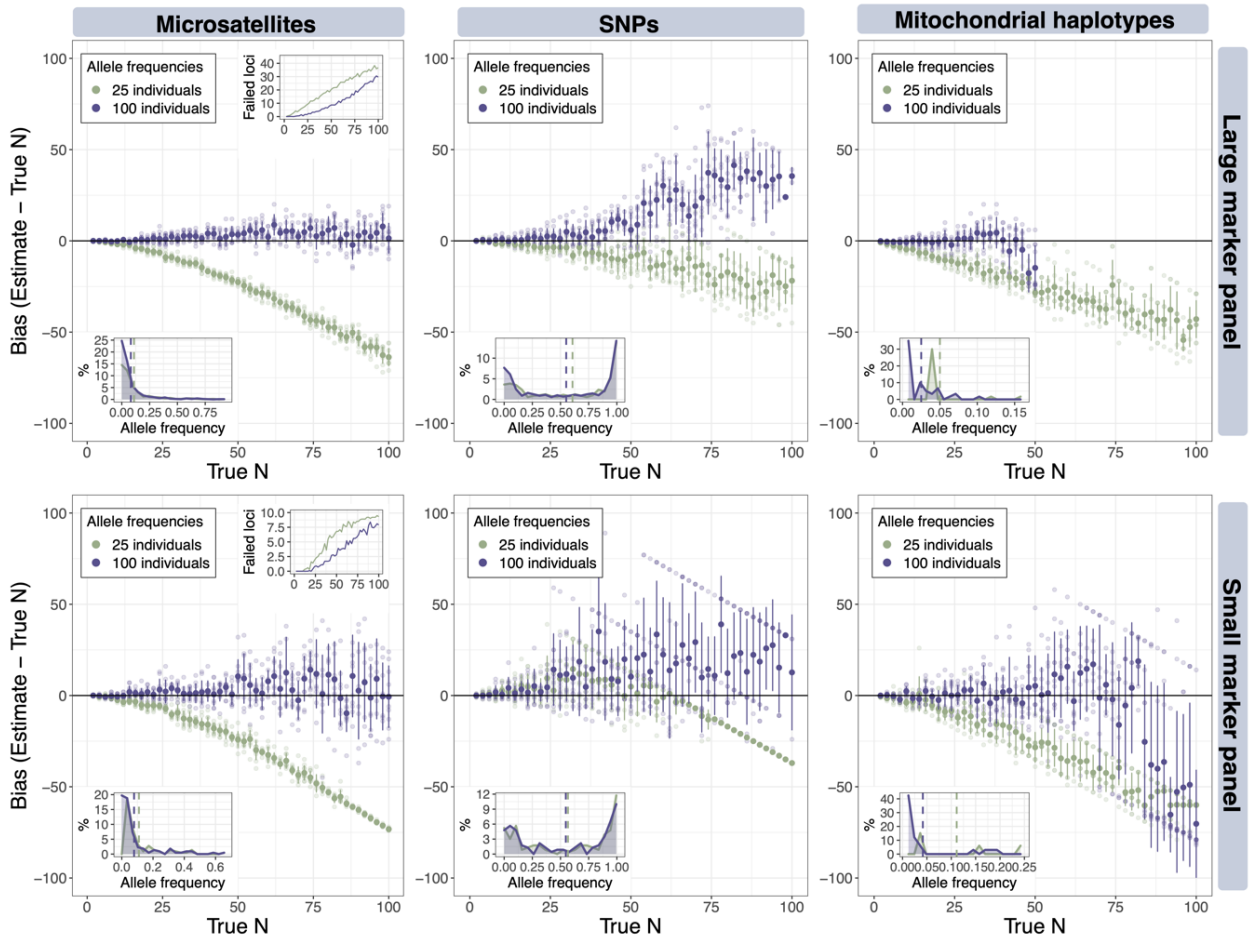

Figure 2.3. Bias in the estimated number of contributors from mixtures of DNA containing genotypes from 2–100 sampled individuals (True N) using large (top row) and small (bottom row) marker panels. Transparent circles show individual DNA mixture simulation results, and solid points and lines show the mean bias ±1 SD across 10 simulated mixtures for each True N. Mixtures were generated for multi-allelic microsatellite markers (left column), bi-allelic

SNP markers (middle column), and single-locus mitochondrial haplotypes (right column). For each marker type, population allele frequencies were calculated using either 25 (purple) or 100 individuals (green) from the full dataset. Inset figures illustrate the distribution of allele frequencies calculated from 25 or 100 individuals, with mean allele frequencies denoted with vertical dashed lines. For microsatellite markers, the inset figure in the top right displays the number of loci for which all alleles were observed as true N increases. These saturated loci were removed from the maximum likelihood calculation.

For the first time, we show that the number of genetic contributors can be estimated in complex DNA mixtures of up to 100 individuals, although the accuracy and precision of the estimate varied depending on marker type, size of the marker panel, and frequency of the rarest allele. Across all three genetic markers, the accuracy of the estimated number of genetic contributors to DNA mixtures is greatest when allele frequencies are estimated using 100 individuals compared to 25 individuals (Table 2.1; Figure 2.3), corresponding to higher resolution allele frequency estimates that can accommodate informative rare alleles. When fewer rare alleles are present in the population and therefore in the DNA mixtures, the model systematically underestimates the number of contributors to mixtures, a problem that worsens as True *N* increases. The size of the marker panel does not have a strong impact on the average bias of the contributor estimation, although the precision of the estimate is greater when more loci or longer mtDNA fragments are used (top panel in Figure 2.3).

All marker types could resolve mixtures of up to 25 contributors, which may be sufficient for many ecological applications. But, to approach larger mixtures, the choice of marker type has an influence the performance of the DNA mixture model. Microsatellite

markers show the greatest promise for accurately estimating the number of individuals in very large DNA mixtures, with average estimates within 2 individuals when True $N = 100$ (Table 2.1; top left panel in Figure 2.3). With allelic richness of up to 20 alleles per locus, both 44-locus and 10-locus marker panels could estimate up to 100 individuals in a DNA mixture, although the number of saturated loci steadily increases as True N increases. While large panels of SNPs can accurately resolve mixtures of up to 50 individuals, the impact of monotonically increasing loci (i.e., loci in which both alleles are observed in the mixture) is apparent in larger mixtures, resulting in an upward bias in the estimated number of individuals (top middle panel). A single mitochondrial fragment of 4 kB contains enough variation to provide accurate estimates of up to 40 individuals, although mixtures larger than this exhibit high numbers of haplotypes and the DNA mixture approach becomes computationally inefficient (top right panel). A smaller mitochondrial fragment can also resolve DNA mixtures of up to 40 individuals, although beyond this point the precision of the estimate is greatly reduced (lower right panel).

Using DNA mixture models in future eDNA research

Efforts to correlate eDNA concentration with species abundance have proven to be moderately successful in providing an index of relative species abundance across space and time. However, an alternative approach to estimating species abundance using the amount of intraspecific genetic diversity detected in eDNA samples may account for variation in eDNA production rate within and among individuals, possibly leading to more accurate estimates of local abundance. We find that DNA mixture models can estimate the number of individuals contributing to mixtures of DNA of up to 100 individuals, provided enough genetic diversity

and rare alleles are present in the mixtures. As reported in Sethi et al. (2019), we found that the presence of rare alleles in a DNA mixture was more important than the number of loci for accurately estimating the number of individuals. Microsatellites or mitochondrial markers are the best marker choices if the expected number of individuals in a mixture of DNA is high because the per-locus information content is much higher than that of bi-allelic SNPs.

Although the simulations described here are important for understanding the limitations and biases in the DNA mixture model, additional research and considerations may be required when using this approach to estimate abundance in eDNA samples. Importantly, our simulated mixtures combined genotypes of sampled individuals with perfect detection of all genetic variants in a mixture. However, imperfect detection of alleles is expected in eDNA samples, particularly when targeting nuDNA markers, and the non-detection of alleles may result in the downward bias of contributor estimates (Sethi et al., 2019). Maximizing the detection of rare alleles may be possible through the improvement of eDNA collection and laboratory approaches including collecting larger volumes of water, target enrichment, or the development of extraction, amplification, and sequencing protocols designed to maximize the recovery of all genetic variation in eDNA samples. Statistical frameworks designed to account for imperfect species detections may also be implemented to account for the imperfect detection of alleles, with models able to evaluate eDNA capture probability, detection probability, and occupancy probability under different environmental conditions (Burian et al., 2021).

An important consideration with mitochondrial markers is the existence of heteroplasmy, where an individual exhibits more than one mtDNA variant. In instances where the minor alleles of heteroplasmic mitochondrial markers are detectable in eDNA, the number

of contributors in mixed DNA samples may be overestimated (Nakanishi et al., 2020). Similarly, any false alleles introduced through PCR error, sequencing error, non-target amplification, or sample contamination will artificially inflate the estimated number of individuals in a mixture. Because rare alleles provide strong evidence about the presence of many individuals in a mixture, the ability to discern between true low-frequency alleles and spurious sequences becomes increasingly important as the true number of contributors increases. Other frameworks developed in criminal forensics research have been able to incorporate information about the magnitude of observed alleles in a mixture (e.g., read counts) as well as model errors such as PCR stutter to improve the interpretation of DNA mixtures (Paoletti, Krane, Doom, & Raymer, 2011; Swaminathan, Grgicak, Medard, & Lun, 2015). Such approaches may be also useful for improving the detection of rare alleles in eDNA mixtures.

Given the results of the simulated DNA mixtures presented here, we provide recommendations for future research aiming to estimate the number of individuals contributing to eDNA samples:

1) Large panels of highly multi-allelic microsatellite markers are the best marker choices for resolving large DNA mixtures. Variable mitochondrial fragments may also be preferable in mixtures of up to 40 individuals, but the calculation becomes impractical when large numbers of haplotypes are observed in large mixtures of DNA. Bi-allelic SNPs do not have the power to estimate the number of individuals contributing to mixtures of DNA containing >25 individuals.
2) To resolve even larger DNA mixtures, genetic markers with a greater number of alleles, and therefore a greater number of rare alleles, will be required.

3) Rare (low frequency) alleles are extremely important for providing information about the presence of many individuals in a mixture of DNA. If large numbers of individuals are expected in a mixture, population allele frequencies should be estimated using as many individuals as possible to increase the number of rare alleles that may be detected. Effort should also be allocated to maximizing the detection of rare alleles in eDNA samples.
4) We recommend that anyone wishing to employ DNA mixture models to estimate abundance from eDNA use a simulation approach as described here to understand the biases and limitations associated with the specific marker panel employed.

Conclusion

The field of eDNA has made substantial progress in recent years, yet environmental samples contain much more genetic information than is currently analyzed in most eDNA research. Recent analyses have focused on intraspecific genetic variation of haplotype diversity from short mitochondrial markers from environmental samples. However, in order to conduct detailed population genetic analyses and estimate species abundance from eDNA, higher-resolution information contained in full mitogenomes or multiple nuclear genetic markers may be required. Although the analysis of intraspecific genetic variation in eDNA samples may introduce additional challenges over species detections, further research on the properties of different genetic markers and the development of bioinformatic and analytical pipelines may allow for robust community-wide population genetic analyses from eDNA.

Acknowledgements

Funding for this project was provided by the University of Michigan Cooperative Institute for Great Lakes Research (CIGLR) through the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement NA17OAR4320152, the Department of Defense (DoD) Strategic Environmental Research and Development Program (SERDP) (grant number RC19-1004), and the Cornell Atkinson Center for Sustainability. We thank S.M. Bogdanowicz, P. Buston, and C. D'Aloia for providing the genetic dataset.

References

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: can environmental DNA (eDNA) cut it as a population genetics tool? *Genes, 10*(3), 192.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology, 21*(1), 1-16.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., . . . Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems, 2*(2), e00191-00116.

Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular Ecology, 30*(3), 685-697.

Anne, C. (2006). Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica, 127*(1), 101-120.

Azarian, C., Foster, S., Devloo-Delva, F., & Feutry, P. (2021). Population differentiation from environmental DNA: Investigating the potential of haplotype presence/absence-based analysis of molecular variance. *Environmental DNA, 3*(3), 541-552.

Baetscher, D. S., Clemento, A. J., Ng, T. C., Anderson, E. C., & Garza, J. C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources, 18*(2), 296-305.

Baker, C. S., Steel, D., Nieukirk, S., & Klinck, H. (2018). Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. *Frontiers in Marine Science, 5*, 133.

Baldigo, B. P., Sporn, L. A., George, S. D., & Ball, J. A. (2017). Efficacy of environmental DNA to detect and quantify brook trout populations in headwater streams of the Adirondack Mountains, New York. *Transactions of the American Fisheries Society, 146*(1), 99-111.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular ecology, 13*(4), 729-744.

Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution, 12*(5), 794-804.

Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics, 17*(1), 1-17. doi:10.1007/s10592-015-0775-4

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation, 29*(7), 2089-2121.

Berner, D. (2019). Allele frequency difference AFD–an intuitive alternative to FST for quantifying genetic population differentiation. *Genes, 10*(4), 308.

Booton, G. C., Kaufman, L., Chandler, M., Oguto-Ohwayo, R., Duan, W., & Fuerst, P. A. (1999). Evolution of the ribosomal RNA internal transcribed spacer one (ITS-1) in

cichlid fishes of the Lake Victoria region. *Molecular Phylogenetics and Evolution, 11*(2), 273-282.

Burian, A., Mauvisseau, Q., Bulling, M., Domisch, S., Qian, S., & Sweet, M. (2021). Improving the reliability of eDNA data interpretation. *Molecular Ecology Resources, 21*(5), 1422-1433.

Bylemans, J., Furlan, E. M., Gleeson, D. M., Hardy, C. M., & Duncan, R. P. (2018). Does size matter? An experimental evaluation of the relative abundance and decay rates of aquatic environmental DNA. *Environmental Science & Technology, 52*(11), 6408-6416.

Bylemans, J., Furlan, E. M., Hardy, C. M., McGuffie, P., Lintermans, M., & Gleeson, D. M. (2017). An environmental DNA-based method for monitoring spawning activity: A case study, using the endangered Macquarie perch (*Macquaria australasica*). *Methods in Ecology and Evolution, 8*(5), 646-655.

Bylemans, J., Gleeson, D. M., Hardy, C. M., & Furlan, E. (2018). Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray–Darling Basin (Australia). *Ecology and Evolution, 8*(17), 8697-8712.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583.

Calus, S. T., Ijaz, U. Z., & Pinto, A. J. (2018). NanoAmpli-Seq: a workflow for amplicon sequencing for mixed microbial communities on the nanopore sequencing platform. *Gigascience, 7*(12), giy140.

Carraro, L., Hartikainen, H., Jokela, J., Bertuzzo, E., & Rinaldo, A. (2018). Estimating species distribution and abundance in river networks using environmental DNA. *Proceedings of the National Academy of Sciences, 115*(46), 11724-11729.

Carreon-Martinez, L. B., Wellband, K. W., Johnson, T. B., Ludsin, S. A., & Heath, D. D. (2014). Novel molecular approach demonstrates that turbid river plumes reduce predation mortality on larval fish. *Molecular ecology, 23*(21), 5366-5377.

Cole, L. W. (2016). The evolution of per-cell organelle number. *Frontiers in Cell and Developmental Biology, 4*, 85.

D'Aloia, C. C., Andrés, J. A., Bogdanowicz, S. M., McCune, A. R., Harrison, R. G., & Buston, P. M. (2020). Unraveling hierarchical genetic structure in a marine metapopulation: A comparison of three high-throughput genotyping approaches. *Molecular Ecology, 29*(12), 2189-2203.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., . . . Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution, 8*(12), 1888-1898.

Deiner, K., Walser, J.-C., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation, 183*, 53-63.

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T., & Minamoto, T. (2017). Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology, 62*(1), 30-39.

Dowle, E. J., Pochon, X., C. Banks, J., Shearer, K., & Wood, S. A. (2016). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: A case study using freshwater macroinvertebrates. *Molecular Ecology Resources, 16*(5), 1240-1254.

Dugal, L., Thomas, L., Reinholdt Jensen, M., Sigsgaard, E. E., Simpson, T., Jarman, S., . . . Meekan, M. (2021). Individual haplotyping of whale sharks from seawater environmental DNA. *Molecular Ecology Resources*.

Dunn, N., Priestley, V., Herraiz, A., Arnold, R., & Savolainen, V. (2017). Behavior and season affect crayfish detection and density inference using environmental DNA. *Ecology and Evolution, 7*(19), 7777-7785.

Dysthe, J. C., Franklin, T. W., McKelvey, K. S., Young, M. K., & Schwartz, M. K. (2018). An improved environmental DNA assay for bull trout (*Salvelinus confluentus*) based on the ribosomal internal transcribed spacer I. *PLoS One, 13*(11), e0206851.

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, 081257.

Egeland, T., Dalen, I., & Mostad, P. F. (2003). Estimating the number of contributors to a DNA profile. *International journal of legal medicine, 117*(5), 271-275.

Elnifro, E. M., Ashshi, A. M., Cooper, R. J., & Klapper, P. E. (2000). Multiplex PCR: optimization and application in diagnostic virology. *Clinical microbiology reviews, 13*(4), 559-570.

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., . . . Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources, 16*(1), 29-41.

Farrell, J. A., Whitmore, L., Mashkour, N., Rollinson Ramia, D. R., Thomas, R. S., Eastman, C. B., . . . Duffy, D. J. (2022). Detection and population genomics of sea turtle species via noninvasive environmental DNA analysis of nesting beach sand tracks and oceanic water. *Molecular Ecology Resources*.

Fediajevaite, J., Priestley, V., Arnold, R., & Savolainen, V. (2021). Meta-analysis shows that environmental DNA outperforms traditional surveys, but warrants better reporting standards. *Ecology and evolution, 11*(9), 4803-4815.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters, 4*(4), 423-425. doi:10.1098/rsbl.2008.0118

Furlan, E. M., Davis, J., & Duncan, R. P. (2020). Identifying error and accurately interpreting environmental DNA metabarcoding results: A case study to detect vertebrates at arid zone waterholes. *Molecular Ecology Resources, 20*(5), 1259-1276.

Furtwängler, A., Reiter, E., Neumann, G. U., Siebke, I., Steuri, N., Hafner, A., . . . Krause, J. (2018). Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. *Scientific Reports, 8*(1), 1-8.

Gantz, C. A., Renshaw, M. A., Erickson, D., Lodge, D. M., & Egan, S. P. (2018). Environmental DNA detection of aquatic invasive plants in lab mesocosm and natural field conditions. *Biological Invasions, 20*(9), 2535-2552.

Gautier, M., Foucaud, J., Gharbi, K., Cezard, T., Galan, M., Loiseau, A., . . . Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology, 22*(14), 3766-3779. doi:10.1111/mec.12360

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., . . . Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351.

Gorički, Š., Stanković, D., Snoj, A., Kuntner, M., Jeffery, W. R., Trontelj, P., . . . Aljančič, G. (2017). Environmental DNA in subterranean biology: range extension and taxonomic implications for *Proteus*. *Scientific Reports, 7*(1), 1-11.

Haned, H., Pene, L., Lobry, J. R., Dufour, A. B., & Pontier, D. (2011). Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? *Journal of Forensic Sciences, 56*(1), 23-28.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1-8.

Hebert, P. D., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 270*(1512), 313-321.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from Pool-seq data. *Genetics, 210*(1), 315-330.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature methods, 12*(4), 351-356.

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., . . . Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources, 21*(3), 690-702.

Jerde, C. L., Mahon, A. R., Chadderton, W. L., & Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters, 4*(2), 150-157. doi:10.1111/j.1755-263X.2010.00158.x

Jo, T., Arimoto, M., Murakami, H., Masuda, R., & Minamoto, T. (2019). Particle size distribution of environmental DNA from the nuclei of marine fish. *Environmental science & technology, 53*(16), 9947-9956.

Jo, T., Arimoto, M., Murakami, H., Masuda, R., & Minamoto, T. (2020). Estimating shedding and decay rates of environmental nuclear DNA with relation to water temperature and biomass. *Environmental DNA, 2*(2), 140-151.

Jo, T., Murakami, H., Masuda, R., Sakata, M. K., Yamamoto, S., & Minamoto, T. (2017). Rapid degradation of longer DNA fragments enables the improved estimation of distribution and biomass using environmental DNA. *Molecular Ecology Resources, 17*(6), e25-e33.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Scientific Reports, 9*(1), 12133. doi:10.1038/s41598-019-48546-x

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagacé, R., Chang, J., Wootton, S., . . . Kidd, J. R. (2014). Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics, 12*, 215-224.

Klymus, K. E., Richter, C. A., Chapman, D. C., & Paukert, C. (2015). Quantification of eDNA shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix. Biological Conservation, 183*, 77-84.

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics, 27*(24), 3435-3436. doi:10.1093/bioinformatics/btr589

Lacoursière-Roussel, A., Côté, G., Leclerc, V., & Bernatchez, L. (2016). Quantifying relative fish abundance with eDNA: a promising tool for fisheries management. *Journal of Applied Ecology, 53*(4), 1148-1157.

Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A., & Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience, 5*(1), s13742-13016-10140-13747.

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS)—Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNA, 2*(3), 255-260.

Long, E. O., & Dawid, I. B. (1980). Repeated genes in eukaryotes. *Annual review of biochemistry, 49*(1), 727-764.

Macé, B., Hocdé, R., Marques, V., Guerin, P. E., Valentini, A., Arnal, V., . . . Manel, S. (2022). Evaluating bioinformatics pipelines for population-level inference using environmental DNA. *Environmental DNA*.

Marshall, N. T., & Stepien, C. A. (2019). Invasion genetics from eDNA and thousands of larvae: A targeted metabarcoding assay that distinguishes species and population variation of zebra and quagga mussels. *Ecology and evolution, 9*(6), 3515-3538.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS One, 9*(12), e114639.

Miller, C. R., Joyce, P., & Waits, L. P. (2002). Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics, 160*(1), 357-366.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp *Cyprinus carpio*. *Molecular Ecology Resources, 17*(2), 324-333.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., . . . Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society open science, 2*(7), 150088. doi:10.1098/rsos.150088

Moushomi, R., Wilgar, G., Carvalho, G., Creer, S., & Seymour, M. (2019). Environmental DNA size sorting and degradation experiment indicates the state of *Daphnia magna* mitochondrial and nuclear eDNA is subcellular. *Scientific Reports, 9*(1), 1-9.

Nakanishi, H., Fujii, K., Nakahara, H., Mizuno, N., Sekiguchi, K., Yoneyama, K., . . . Saito, K. (2020). Estimation of the number of contributors to mixed samples of DNA by mitochondrial DNA analyses using massively parallel sequencing. *International journal of legal medicine, 134*(1), 101-109.

Olson, Z. H., Briggler, J. T., & Williams, R. N. (2012). An eDNA approach to detect eastern hellbenders (*Cryptobranchus a. alleganiensis*) using samples of water. *Wildlife Research, 39*(7). doi:10.1071/wr12114

Paoletti, D. R., Doom, T. E., Krane, C. M., Raymer, M. L., & Krane, D. E. (2005). Empirical analysis of the STR profiles resulting from conceptual mixtures. *Journal of Forensic Sciences, 50*(6).

Paoletti, D. R., Krane, D. E., Doom, T. E., & Raymer, M. (2011). Inferring the number of contributors to mixed DNA profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(1), 113-122.

Parsons, K. M., Everett, M., Dahlheim, M., & Park, L. (2018). Water, water everywhere: Environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science, 5*(8), 180537.

Piggott, M. P. (2016). Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish. *Ecology and Evolution, 6*(9), 2739-2750.

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., & Waits, L. P. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences, 70*(8), 1123-1130.

Pinfield, R., Dillane, E., Runge, A. K. W., Evans, A., Mirimin, L., Niemann, J., . . . Foote, A. D. (2019). False-negative detections from environmental DNA collected in the presence of large numbers of killer whales (*Orcinus orca*). *Environmental DNA, 1*(4), 316-328.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics, 6*(11), 847-859.

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes, 7*(3), 355-364.

Robin, E. D., & Wong, R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *Journal of cellular physiology, 136*(3), 507-513.

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation, 17*, e00547.

Schmelzle, M. C., & Kinziger, A. P. (2016). Using occupancy modelling to compare environmental DNA to traditional field methods for regional-scale monitoring of an endangered aquatic species. *Molecular Ecology Resources, 16*(4), 895-908.

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources, 15*(6), 1289-1303.

Sepulveda, A. J., Schabacker, J., Smith, S., Al-Chokhachy, R., Luikart, G., & Amish, S. J. (2019). Improved detection of rare, endangered and invasive trout in using a new

large-volume sampling method for eDNA capture. *Environmental DNA, 1*(3), 227-237.

Sethi, S. A., Larson, W., Turnquist, K., Isermann, D., & Kembel, S. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution, 10*(1), 109-119. doi:10.1111/2041-210x.13079

Shogren, A. J., Tank, J. L., Andruszkiewicz, E., Olds, B., Mahon, A. R., Jerde, C. L., & Bolster, D. (2017). Controls on eDNA movement in streams: Transport, Retention, and Resuspension. *Scientific Reports, 7*(1), 5065. doi:10.1038/s41598-017-05223-1

Shum, P., & Palumbi, S. R. (2021). Testing small-scale ecological gradients and intraspecific differentiation for hundreds of kelp forest species using haplotypes from metabarcoding. *Molecular Ecology*.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Moller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA-Current status and future perspectives. *Evolutionary Applications, 13*(2), 245-262. doi:10.1111/eva.12882

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., . . . Thomsen, P. F. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution, 1*(1), 4. doi:10.1038/s41559-016-0004

Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics, 9*(6), 477-485.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*(1), 195-197.

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., . . . Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports, 7*(1), 1-11.

Stepien, C. A., Snyder, M. R., & Elz, A. E. (2019). Invasion genetics of the silver carp *Hypophthalmichthys molitrix* across North America: Differentiation of fronts, introgression, and eDNA metabarcode detection. *PLoS One, 14*(3), e0203012.

Strickler, K. M., Fremier, A. K., & Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation, 183*, 85-92.

Swaminathan, H., Grgicak, C. M., Medard, M., & Lun, D. S. (2015). NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. *Forensic Science International: Genetics, 16*, 172-180. doi:10.1016/j.fsigen.2014.11.010

Székely, D., Corfixen, N. L., Mørch, L. L., Knudsen, S. W., McCarthy, M. L., Teilmann, J., . . . Olsen, M. T. (2021). Environmental DNA captures the genetic diversity of bowhead whales (*Balaena mysticetus*) in West Greenland. *Environmental DNA, 3*(1), 248-260.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., . . . Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research, 24*(16), 3189-3194.

Toju, H., Tanabe, A. S., Yamamoto, S., & Sato, H. (2012). High-coverage ITS primers for the DNA-based identification of ascomycetes and basidiomycetes in environmental samples. *PLoS One, 7*(7), e40863.

Tsuji, S., Maruyama, A., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020). Environmental DNA analysis shows high potential as a tool for estimating intraspecific genetic diversity in a wild fish population. *Molecular Ecology Resources, 20*(5), 1248-1258.

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA, 2*(1), 42-52.

Tsuji, S., Shibata, N., Sawada, H., & Ushio, M. (2020). Quantitative evaluation of intraspecific genetic diversity in a natural fish population using environmental DNA analysis. *Molecular Ecology Resources, 20*(5), 1323-1332.

Turner, C. R., Barnes, M. A., Xu, C. C., Jones, S. E., Jerde, C. L., & Lodge, D. M. (2014). Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods in Ecology and Evolution, 5*(7), 676-684.

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications, 30*(2), e02036.

Uchii, K., Doi, H., & Minamoto, T. (2016). A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Molecular Ecology Resources, 16*(2), 415-422.

Uchii, K., Doi, H., Yamanaka, H., & Minamoto, T. (2017). Distinct seasonal migration patterns of Japanese native and non-native genotypes of common carp estimated by environmental DNA. *Ecology and Evolution, 7*(20), 8515-8522.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., . . . Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Wei, N., Nakajima, F., & Tobino, T. (2018). A microcosm study of surface sediment environmental DNA: decay observation, abundance estimation, and fragment length comparison. *Environmental Science & Technology, 52*(21), 12428-12435.

Weitemier, K., Penaluna, B. E., Hauck, L. L., Longway, L. J., Garcia, T., & Cronn, R. (2021). Estimating the genetic diversity of Pacific salmon and trout using multigene eDNA metabarcoding. *Molecular Ecology, 30*(20), 4970-4990.

Wilcox, T. M., McKelvey, K. S., Young, M. K., Lowe, W. H., & Schwartz, M. K. (2015). Environmental DNA particle size distribution from Brook Trout (*Salvelinus fontinalis*). *Conservation Genetics Resources, 7*(3), 639-641.

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., & Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. *Molecular Ecology Resources, 18*(6), 1392-1401.

Yates, M. C., Fraser, D. J., & Derry, A. M. (2019). Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. *Environmental DNA, 1*(1), 5-13.

Yoccoz, N. G., Nichols, J. D., & Boulinier, T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution, 16*(8), 446-453.

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution, 11*(12), 1609-1625.

CHAPTER 3

NUCLEAR EDNA ESTIMATES POPULATION ALLELE FREQUENCIES AND ABUNDANCE IN EXPERIMENTAL MESOCOSMS AND FIELD SAMPLES[2]

Abstract

Advances in environmental DNA (eDNA) methodologies have led to improvements in the ability to detect species and communities in aquatic environments, yet the majority of studies emphasize biological diversity at the species level by targeting variable sites within the mitochondrial genome. Here, we demonstrate that eDNA approaches also have the capacity to detect intraspecific diversity in the nuclear genome, allowing for assessments of population-level allele frequencies and estimates of the number of genetic contributors in a sample. Using a panel of microsatellite loci developed for the Round Goby (*Neogobius melanostomus*), we tested the similarity between eDNA-based and individual tissue-based estimates of allele frequencies from experimental mesocosms and in a field-based trial. Subsequently, we used a likelihood-based DNA mixture framework to estimate the number of unique genetic contributors in eDNA samples and in simulated mixtures of alleles. In both mesocosm and field samples, allele frequencies from eDNA were highly correlated with allele frequencies from genotyped Round Goby tissue samples, indicating nuclear markers can be reliably amplified from water samples. DNA mixture analyses were able to estimate the number of

[2] This chapter has been published as Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular ecology, 30*(3), 685-697; it is reprinted here with permission.

genetic contributors from mesocosm eDNA samples and simulated mixtures of DNA from up to 58 individuals, with the degree of positive or negative bias dependent on the filtering scheme of low-frequency alleles. With this study we document the application of eDNA and multiple amplicon-based methods to obtain intraspecific nuclear genetic information and estimate the absolute abundance of a species in eDNA samples. With proper validation, this approach has the potential to advance non-invasive survey methods to characterize populations and detect population-level genetic diversity.

Introduction

Environmental DNA (eDNA) methods are transforming how scientists and resource managers assess the diversity and distributions of organisms (Deiner, Bik, et al., 2017; Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012). Using DNA isolated from environmental samples such as ancient and terrestrial sediments, ice cores, and aquatic ecosystems, eDNA methodologies capture the genetic material organisms release into the environment through cells, hair, skin, and feces (Thomsen et al., 2012; Willerslev et al., 2007; Willerslev et al., 2003). Such approaches can provide an efficient way to detect species presence/absence (Ficetola, Miaud, Pompanon, & Taberlet, 2008; Pilliod, Goldberg, Arkle, & Waits, 2013), habitat use (Stewart, Ma, Zheng, & Zhao, 2017), and relative abundance (Hänfling et al., 2016; Jerde, Mahon, Chadderton, & Lodge, 2011). With greater detection probability and reduced cost over traditional sampling methods, eDNA methods are particularly well-suited for surveillance of aquatic invasive species, where early detection may be vital for their management or eradication (Dejean et al., 2012; Jerde et al., 2011; Lodge et al., 2016; Vander Zanden, Hansen, Higgins, & Kornis, 2010). Furthermore, technical advancements in next-generation sequencing (NGS) methods have led to the development of eDNA metabarcoding, or the simultaneous detection of multiple species with a single molecular marker (e.g. Deiner, Bik, et al., 2017; Kelly, Port, Yamahara, & Crowder, 2014; Margulies et al., 2005; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; Valentini et al., 2016). Environmental DNA can therefore provide information about species distributions, relative abundance, or composition that can be broadly applied in studies of biodiversity, community ecology, and conservation biology (Bohmann et al., 2014; Lodge et al., 2012; Thomsen & Willerslev, 2015).

The majority of eDNA studies to date have assessed biological diversity at or above the species level, with relatively little attention given to intraspecific genetic diversity (Adams et al., 2019; Sigsgaard et al., 2020). However, some recent studies have developed approaches to detect intraspecific genetic variation in the mitochondrial genome from environmental samples (Deiner, Renshaw, et al., 2017; Elbrecht, Vamos, Steinke, and Leese, 2018; Parsons, Everett, Dahlheim, & Park, 2018; Sigsgaard et al., 2017; Tsuji et al., 2019; Turon, Antich, Palacín, Præbel, & Wangensteen, 2020). Due to its high copy number per cell, mitochondrial DNA (mtDNA) may occur at higher concentrations in water than nuclear DNA (but see Bylemans et al., 2017; Minamoto et al., 2017; Piggott, 2016), potentially leading to higher detection probability in environmental samples. On the other hand, mtDNA is haploid and non-recombining so, as a single locus, may be limited in providing the high resolution of genetic variation required for detailed population genetic analyses (Ballard & Whitlock, 2004; Hurst & Jiggins, 2005; Rubinoff, Cameron, & Will, 2006; Teske et al., 2018). Expanding eDNA approaches to detect intraspecific variation in nuclear DNA markers such as microsatellites or single nucleotide polymorphisms (SNPs) can therefore enhance our ability to make genetic inferences at the population level.

In this study, we explore the extent to which intraspecific genetic diversity can be detected in eDNA and used to estimate the number of unique genetic contributors to an eDNA sample. As a proof of concept, we use nuclear microsatellite markers and NGS methods to characterize population allele frequencies and estimate the absolute abundance of Round Gobies (*Neogobius melanostomus*) using eDNA samples from experimental mesocosms and in a field-based trial. The Round Goby, a fish species native to the Ponto-Caspian region, was initially introduced to North America via ballast water in 1990 and has since spread

throughout the Laurentian Great Lakes (Charlebois et al., 1997; Jude, Reider, & Smith, 1992; Schaeffer, Bowen, Thomas, French III, & Curtis, 2005). More recently, Round Gobies have spread to inland lakes and rivers, where they can cause native species declines through competition, predation, and contaminant cycling (Janssen & Jude, 2001; Kornis, Mercado-Silva, & Vander Zanden, 2012; Krakowiak & Pennuto, 2008). Due to the short time interval between arrival and establishment, Round Gobies present a high invasion risk even at low densities (Vélez-Espino, Koops, & Balshine, 2010), and control strategies may require information on species abundance due to the rapid decline in the success of eradication efforts as invasive populations grow and spread (Vander Zanden et al., 2010). Thus, the development of eDNA methods to quantify species abundance at the invasion front could lead to improved management strategies for this invader.

Several previous efforts to assess species abundance with eDNA have used correlative relationships between eDNA concentration and indices of species abundance or biomass (e.g. Kelly et al., 2014; Pilliod et al., 2013; Takahara, Minamoto, Yamanaka, Doi, & Kawabata, 2012). While these methods can provide an index of relative abundance, their accuracy and precision with respect to absolute species abundance has been difficult to establish, and such correlative relationships can be heavily impacted by taxon-specific amplification biases (Kelly, Shelton, & Gallego, 2019) or local biotic and abiotic factors influencing the amount of DNA shed by an organism (Barnes & Turner, 2015). For instance, the production rate of eDNA can vary with an organism's size, behavior, or metabolism, all of which may vary across a range of abiotic conditions (Klymus, Richter, Chapman, & Paukert, 2015; Lacoursière-Roussel, Rosabal, & Bernatchez, 2016; Maruyama, Nakamura, Yamanaka, Kondoh, & Minamoto, 2014; Takahara et al., 2012). The difficulty in obtaining robust

quantitative measurements of eDNA production among individuals and its relationship to the amount of DNA in an eDNA sample currently limits our ability to reliably link measurements of eDNA concentration to species abundance, density, or biomass (Iversen, Kielgast, & Sand-Jensen, 2015).

In contrast to correlative relationships between eDNA concentrations and relative species abundance, DNA mixture estimators take a radically different approach to estimating absolute abundance in a sample (Sethi, Larson, Turnquist, & Isermann, 2019). Originally developed in criminal forensics, DNA mixture estimators provide an inferential framework that uses the genetic signature of mixtures to estimate the number of unique genetic contributors in a mixture of DNA based on population allele frequencies and the number of unique alleles identified (Curran, Triggs, Buckleton, & Weir, 1999; Weir et al., 1997). While these estimators have previously been applied to tissue-based mixtures of DNA for diet analysis (Sethi et al., 2019), environmental samples can also contain DNA from multiple individuals. If intraspecific genetic diversity can be detected eDNA, mixture estimators may therefore provide a means of estimating the number of contributors to environmental samples that relies on the detected presence of haplotypes or alleles rather than eDNA concentrations.

Here, we apply DNA mixture estimators to eDNA using species-specific nuclear genetic markers we developed for Round Gobies. We first assess the similarity of allele frequencies from eDNA and individually genotyped individuals in experimental mesocosms to evaluate the extent to which alleles derived from Round Goby tissues are represented in sequence data recovered from eDNA. We then use a likelihood-based DNA mixture model to estimate the number of genetically unique individuals contributing genetic material to each eDNA sample. Finally, we test the ability of the DNA mixture estimator to accurately

estimate the number of unique genetic contributors in simulated combinations of up to 58 individuals.

Materials and methods

Microsatellite characterization and multiplex assay development

Genomic DNA (50-100 ng) from a pool of three Round Goby (*Neogobius melanostomus*) individuals collected from Cayuga Lake, New York, USA was endonuclease-digested with *AluI*, *RsaI*, and *Hpy166II*. The digestions were pooled for subsequent adenylation with Klenow (exo-) and dATP, and the resulting products were ligated to an Illumina Y-adaptor sequence using T4 DNA ligase in the presence of 1 mM ATP. Genomic fragments containing repeats were captured by hybridization to biotinylated repeats and streptavidin-coated magnetic beads followed by amplification with Platinum Taq DNA polymerase and indexing with Illumina primers (one universal primer and one index primer). PCR products were quantified with a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA), verified by electrophoresis on a 1.0% agarose gel, and size-selected (300–600 bp) with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). The 'design primers' function of MSATCOMMANDER 1.0.3 (Faircloth, 2008; Rozen & Skaletsky, 2000) was then used to create a library of microsatellite tetramer repeats based on the number of motif repeats (10–24) and PCR product length (410–440 bp).

Primer specificity was inspected using NCBI Primer Blast, where no other species were detected as matches to the designed primer pairs. Forward and reverse primers (range 20–24 bp) for 43 loci were ordered from Integrated DNA Technologies (http://www.idtdna.com) and tested for functionality in single reactions using genomic DNA extracted from the tissue of

three Round Gobies. Following exclusion of primers with complementary sequences or sub-optimal PCR amplification, 35 microsatellite loci remained (Table S3.1).

Microsatellite loci were grouped into seven multiplex PCR assays, with each multiplex containing 4–6 primer pairs (Table S3.1). We tested the performance of each multiplex in a PCR containing 1 µL (20–30 ng) of Round Goby genomic DNA, 1 µL of primer pairs in equimolar concentrations (2 µM), and 5 µL of QIAGEN Multiplex PCR Master Mix (QIAGEN, Inc.). The program for multiplex PCR is as follows: initial denaturation at 95°C for 15 minutes followed by 35 cycles of 94° C for 30 s, 59° C for 90 s, and 72°C for 90 s. Gel electrophoresis in 1% agarose stained with ethidium bromide confirmed the presence of PCR products within the expected size range for all multiplexes.

Mesocosm experiment

We collected live Round Gobies (n = 58) from a site on Cayuga Lake via beach seining and placed them in one of 12 experimental mesocosms containing 12 L of aged room temperature tap water. Each mesocosm treatment was conducted in triplicate and contained Round Gobies (approximately 7–12 cm length) at densities of one, three, five, or ten individuals. An additional Round Goby was erroneously added to a single replicate of the n = 10 treatment to total 11 individuals (labeled mesocosm 10c), but is hereafter grouped into the density treatment of ten individuals. Two additional mesocosms served as negative controls (mesocosms with aged room temperature tap water only). After one hour, Round Gobies were removed from the mesocosms and euthanized with MS-222 according to the Cornell IACUC Animal Care and Use Procedure (ACUP 306.02). Tissues were sampled from caudal fins of each individual and DNA was extracted with a DNeasy Blood and Tissue extraction kit

(QIAGEN, Inc.) following manufacturer protocols. Following the removal of all fish from the mesocosms, duplicate 2-L water samples were collected from each mesocosm in sterilized wide-mouth Nalgene™ plastic bottles and stored on ice until vacuum-filtration through a cellulose nitrate membrane filter (47 mm diameter, 1 μm pore size). Filters were immersed in 700 μL Longmire's solution (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5 % SDS) and stored at -20° C until DNA extraction. Environmental DNA was extracted from filters following a modified protocol from the DNeasy Blood and Tissue extraction kit (QIAGEN, Inc.) as in Spens et al. (2017). To minimize contamination, eDNA sample filtration and pre-PCR laboratory protocols were carried out in separate rooms within dedicated pre-PCR facilities, and stringent precautions were followed according to Goldberg et al. (2016). Round Goby tissues were handled and processed in a separate facility. All reusable equipment including collection bottles, forceps, and the vacuum filtration apparatus was cleaned between samples by soaking in a 50% commercial bleach solution, rinsing in DI water, and treating under UV bulbs for 30 minutes each. In addition to the two field controls described above, one filtration blank and one PCR blank served as negative controls.

Field trial

To determine the feasibility of estimating population allele frequencies from eDNA samples in a field-based setting, we collected eDNA samples and additional Round Gobies (n = 15) from another site on Cayuga Lake (c. 20 miles away from the site of Round Goby collection for the mesocosm experiment; Figure S3.1A). Sampling was conducted during the summer months when Round Goby densities peak in nearshore waters, and density estimates from a previous study using benthic videography and direct observation report Round Goby

densities of 0.34 fish/m^2 in this section of the lake (Andres et al., 2020). We confirmed the Round Gobies collected from the two sites in Cayuga Lake are panmictic using genotyped tissue samples from both sites and the 'find.clusters' function of the ADEGENET package in R version 3.5 (Jombart, 2008; R Core Team, 2016), where a single cluster (k) exhibited the lowest Bayesian Information Criterion (BIC; Figure S3.1B). Thus, we consider all 73 Round Gobies (58 in the mesocosm experiment and 15 in the field trial) when estimating tissue-based population allele frequencies in the field trial. To sample eDNA, three 2-L water samples were collected from shoreline locations approximately 50 m apart in sterilized wide-mouth Nalgene™ plastic bottles. A negative field control of 2-L of distilled water was also collected at the site. Water filtration, tissue sampling, and DNA extraction protocols were identical to those described for the mesocosm experiment above.

Library preparation and MiSeq sequencing

Microsatellite loci were amplified from eDNA and tissue samples in separate reactions using multiplex PCR methods described above, with the number of PCR cycles increased to 45 for eDNA samples due to low template DNA concentrations. Three PCR replicates were performed for each of the three eDNA samples from the field trial. Products from all seven multiplexes were pooled from each sample in equal volumes (5 µL each) and uniquely barcoded in a second-stage PCR using Illumina Nextera XT tags. Each 20-µL second-stage PCR included 2 µL pooled PCR product diluted 1:1 with molecular H_2O, 4 µL 5x HF buffer, 0.4 µL 10 mM dNTPs, 0.1 µL OneTaq DNA polymerase, 0.4 µL each of 10 µM Nextera Index Primer 1 (N701–N728) and Nextera Index Primer 2 (N502–N521). One library was constructed from the pooled PCR products for all tissue and eDNA samples in the mesocosm

experiment, while another library was constructed from the tissue and eDNA samples from the field trial. DNA libraries were purified with Agencourt AMPure XP beads and the concentration of each library was estimated using the Qubit dsDNA High-Sensitivity Kit and Qubit 2.0 fluorometer. The libraries were diluted to 4 nM with PCR-grade water and paired-end sequenced on an Illumina MiSeq sequencing platform (Illumina, San Diego, CA) with the MiSeq v2 500 bp kit (PE 2 x 250 bp) by Cornell University's Institute of Biotechnology Genomics Facility.

Bioinformatic analysis

Demultiplexed reads from each Miseq run were processed with TRIMMOMATIC v0.39 (Bolger, Lohse, & Usadel, 2014) to remove adapter sequences. We then ran a custom Perl script to extract forward and reverse reads and assign them to each locus as described in D'Aloia, Bogdanowicz, Harrison, & Buston (2017). The script includes the following steps: 1) trim low-quality reads with Phred scores less than 20; 2) create contigs from overlapping paired-end reads with a minimum overlap of at least 20 bp and mismatch rate of less than 0.05; 3) identify and sort reads corresponding to each locus using the forward primer; 4) collapse identical reads (100% identity) for each sample; and 5) collapse reads across all samples. To filter out most PCR artifacts and paralogs while retaining true microsatellite repeats and SNPs, we required 90% of the first 40 base pairs of a read to align with and match the reference contig constructed from the most common allele at each locus across all of the samples. We determined the multilocus diploid genotype for each Round Goby tissue sample based on the allele with the highest read count at each locus. Individuals were considered heterozygous at a locus if at least 20% of the reads corresponded to a second allele, and only

alleles with a read depth of at least 10 reads per individual were considered (as in D'Aloia et al., 2017). Following individual genotyping, we excluded two poorly amplified loci and five potentially paralogous loci exhibiting significant deviations from Hardy-Weinberg equilibrium (Paradis, 2010) and heterozygote excess. The remaining 28 loci were used in further analyses (Table S3.1).

For eDNA samples, we excluded alleles with fewer than ten total reads in each sample and scaled read counts to 100 reads per sample to account for differences in read depth. To further filter out potentially erroneous sequence data arising from PCR stutter and sequencing error, we removed alleles below 1% frequency in each eDNA sample from analysis. Due to low variation in read depth and allele frequencies between duplicate mesocosm eDNA samples (Figures S3.2–S3.3), we pooled the scaled reads from the two eDNA samples for each mesocosm. We also pooled the scaled reads from the three replicate eDNA samples from the field trial. eDNA allele frequencies were then estimated as the read frequencies of alleles in each mesocosm and in the field eDNA sample. Thus, while allele frequency estimations in tissue samples are derived from genotyped individuals, allele frequency estimations in eDNA samples are taken directly from sequence read frequencies.

Comparison of genotyped individuals and eDNA samples

All further analyses were performed in R version 3.5 (R Core Team, 2016). To determine the similarity between allele frequencies derived from eDNA reads and genotyped tissues in the mesocosm experiment, we combined allele frequencies across all 12 mesocosm eDNA samples and evaluated the correlation between eDNA allele frequencies and tissue allele frequencies for all alleles across all loci, as well as on a per-locus basis. We further

examined the similarity between eDNA-based and tissue-based allele frequencies in corresponding mesocosms by conducting a principal components (PC) analysis on the scaled and centered allele frequencies from eDNA reads and genotyped individuals. Subsequently, we constructed a Euclidean distance matrix for all samples using principal components values along all PC axes described above as inputs. For the field trial, we evaluated the correlation between allele frequencies determined from the eDNA samples collected from Cayuga Lake and from the 73 genotyped Round Gobies.

DNA mixture contributor estimation

To estimate the number of unique genetic contributors to a DNA mixture (e.g. the number of individuals captured in each eDNA sample), we implemented a likelihood-based model described in Sethi et al. (2019). At each locus *j*, the model estimates the likelihood that a proposed number of diploid contributors, *x*, produces the observed set of *n* alleles, $A = \{a_1, \ldots, a_n\}$, given a set of associated population allele frequencies, $p = \{p_1, \ldots, p_n\}$, using the following equation:

$$L_j(x|A,p) = \sum_{d_1=0}^{d} \sum_{d_2=0}^{d-d_1} \cdots \sum_{d_{n-1}}^{d-d_1-\ldots-d_{n-2}} \left[\left(\frac{(2x)!}{\prod_{i=1}^{n} g_i!} \right) \prod_{i=1}^{n} p_i^{g_i} \right] \quad (1)$$

This equation accounts for all of the combinations of alleles that may arise in a mixture due to redundancy within or among individuals, where $d = 2x - n$ is the total number of 'masked' alleles calculated as the difference between the total number of alleles present for x diploid organisms and the total number of unique alleles observed in the mixture

genotype, and g_i defined the total number of copies of allele a_i truly present in the mixture plus any masked copies of the allele d_i, with $\sum_{i=1}^{n} g_i = 2x$. As in Sethi et al. (2019), we calculated this likelihood with custom R scripts using a numerically equivalent but more computationally efficient form of Eq. 1 derived by Weir et al. (1997).

The estimated number of individuals contributing to the DNA mixture is therefore identified as the maximum likelihood estimate of the number of contributors given the product of this likelihood across all loci:

$$\max_{x} \prod_{j} L_j(x|A, p) \tag{2}$$

For the mesocosm samples, we applied this equation to pooled individual genotypes and eDNA mixtures from each mesocosm with a proposed number of contributors (1–100), where the set of observed alleles A was determined per mesocosm and the population allele frequencies *p* were estimated directly from the 58 genotyped individuals used in the experiment. To evaluate the sensitivity of the contributor estimation to false alleles and allelic dropout, we filtered eDNA sequence reads according to a succession of increasingly strict thresholds, or frequencies below which reads were removed (0.001, 0.01, 0.1). Due to variation in the number of alleles present at each locus (Table S3.1), we also filtered reads using variable thresholds according to per-locus allelic richness, where the threshold decreased from 0.1 to 0.001 as the number of alleles at a locus increased. We repeated the contributor estimations using the allele frequencies combined across all eDNA samples to represent population allele frequencies *p*. Bias in the contributor estimation (estimated # contributors - true # contributors) was calculated for each eDNA-based and tissue-based DNA mixture.

To assess the performance of the contributor estimation on eDNA samples representing a greater number of individuals, we applied the maximum likelihood estimator to simulated mixtures of up to our total sample of 58 Round Gobies in the mesocosm experiment. Using a bootstrapping procedure, we combined eDNA read counts from mesocosms in simulated mixtures ranging from 2–12 mesocosms per draw. We estimated the number of genetic contributors to mixtures with 1,000 bootstrap replicates at fixed thresholds and a variable threshold based on allelic richness as described above.

We also applied the contributor estimation to each eDNA sample from the field trial, where the set of observed alleles A was determined from each eDNA sample and population allele frequencies *p* were estimated from the 73 genotyped individuals used in the experiment. We repeated the contributor estimations with allele frequencies combined across the three replicate eDNA samples taken from the field used to represent population-level allele frequencies *p*.

Results

Sequencing and genotyping

The full dataset contained 47,920,390 reads, of which 35,583,440 remained after demultiplexing and trimming adapters. Following exclusion of alleles below the minimum read depth of 10 reads, the target loci were not identified in any of the negative control blanks from the field, extraction, or amplification processes, indicating there was no detectable cross-contamination. Round Goby tissue samples exhibited a high total read depth per sample (mean = 45,534 reads, s.d. = 19,958; Figure S3.4A) and total read depth per locus (mean = 1,626 reads, s.d. = 1,714, Figure S3.4B). All individuals were genotyped at ≥ 26 of the 28 loci

in all individuals (i.e. fewer than two loci per sample were considered missing data in our pipeline). All microsatellites were multiallelic with an average of 9.4 alleles per locus (range: 2–21 alleles per locus) among the 73 Round Gobies comprising the sample population (Table S3.1). Microsatellite loci were successfully amplified in all eDNA samples from mesocosms containing fish with an average total read depth of 37,151 reads per sample (s.d. = 9,161) and average read depth of 1,327 reads per locus (s.d. = 1,393). The average per-locus read depth and total read depth did not vary across mesocosm densities (Figure S3.5), indicating Round Goby density did not have an impact on sequence recovery in the mesocosm experiment. Read depths were lower in eDNA samples from the field trial, with an average total read depth of 4,305 reads per sample (s.d. = 3,796; Figure S3.4A) and 154 reads per locus (s.d. = 283; Figure S3.4B).

Comparison of genotyped individuals and eDNA samples

In the mesocosm experiment, allele frequencies from eDNA sequence reads across all mesocosms closely resemble allele frequencies from the 58 genotyped individuals (Pearson's correlation coefficient r = 0.95 across all loci, range r = 0.88–1.00 per locus; Figure 3.1A). Principal component analysis results showed high similarity between eDNA samples and genotyped individuals in each mesocosm, where eDNA samples clustered tightly with the individuals from the associated mesocosm (Figure 3.1C). Across all PC axes, the pairwise Euclidean distance was noticeably smaller within a mesocosm than between mesocosms (Figure 3.1D). eDNA and tissue sample pairs were most differentiated from other samples when derived from mesocosms with single Round Gobies, becoming more genetically similar to other mesocosms as the number of Round Gobies per mesocosm increased.

Allele frequencies from eDNA samples and genotyped individuals were also highly correlated in the field trial, with a Pearson's correlation coefficient of r = 0.84 across all loci (range r = 0.41–1.00 per locus; Figure 3.1B). Several alleles at low frequency in the population were not recovered by eDNA samples, with only 121 of 253 total alleles identified from genotyped individuals occurring in at least one of the three eDNA samples. However, all alleles with a frequency > 0.24 in the 73 genotyped individuals were recovered by at least one of the three eDNA samples, and alleles not detected with eDNA occurred at low frequencies in the population (mean = 0.03, s.d. = 0.04). On the other hand, eDNA samples also identified several alleles not documented in the genotyped individuals, albeit at low read frequencies (mean = 0.02, s.d. = 0.02). Such alleles may represent true low-frequency alleles not included in the genotyped individuals or may be the product of erroneous sequences.

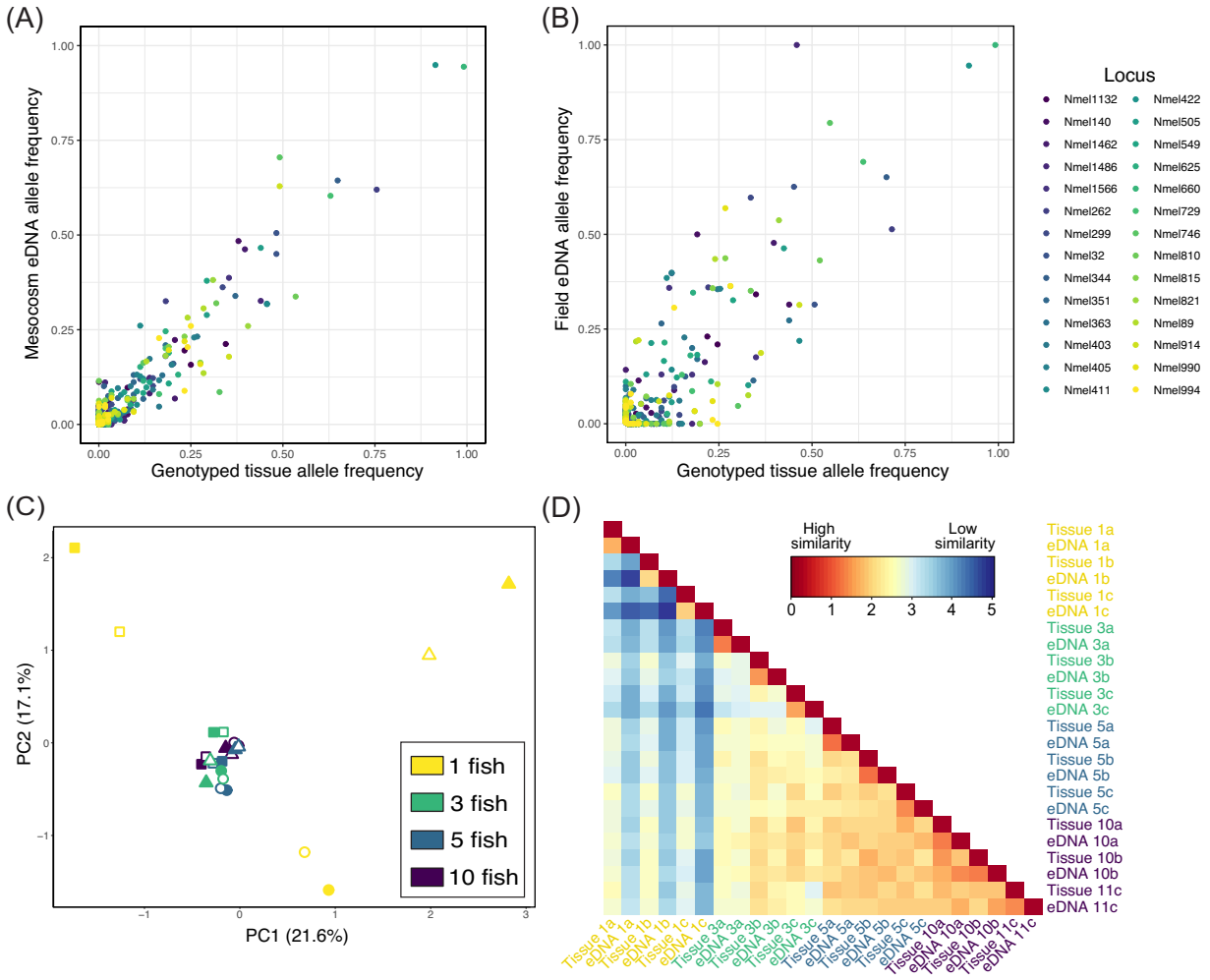

Figure 3.1. (A) Correlation between eDNA-derived and tissue-derived allele frequencies for all alleles across 28 loci in the mesocosm experiment. (B) Correlation between eDNA-derived and tissue-derived allele frequencies for all alleles across 28 loci in the field trial. (C) PCA of allele frequencies across 28 loci for Round Goby tissue samples (filled symbols) and eDNA samples (hollow symbols) from 12 mesocosms varying in Round Goby density. Colors represent mesocosm density treatments (1, 3, 5, or 10 fish) and symbols represent treatment replicates. (D) Heatmap of the pairwise Euclidean distances across all PC axes of allele frequencies from mesocosm eDNA and tissue samples, with blue colors indicating far

distances (low similarity) and red colors indicating close distances (high similarity). Samples are arranged in pairs (eDNA/tissue samples) from each mesocosm, with colors representing mesocosm density treatments and letters (a, b, or c) representing treatment replicates.

Contributor estimation

Estimates of the number of genetic contributors in mesocosms using observed alleles from genotyped tissue samples were within ±2 contributors at all Round Goby densities when population-level allele frequencies were specified using genotyped tissues and in mixtures of up to 5 individuals when allele frequencies were specified from eDNA read frequencies (top panel, Figure 3.2). When estimating the number of genetic contributors using observed alleles from eDNA samples, patterns of bias emerged across frequency thresholds below which reads were removed (0.001, 0.01, 0.1) regardless of how population-level allele frequencies were characterized. The contributor estimation was positively biased the lowest thresholds (0.001 and 0.01) across all mesocosm densities with the exception of the 10-individual mixtures using population allele frequencies from genotyped individuals, where estimates were within ±1 genetic contributor. Contributor estimations were also within ±1 contributors in mesocosms with one or three Round Gobies at the highest threshold (0.1), while negative bias was more apparent in mesocosms with five or ten individuals at this threshold (Figure 3.2). Across all mesocosm densities, the variable threshold based on allelic richness outperformed all other thresholds (maximum bias = +5 associated with a 10-individual mixture using population allele frequencies from eDNA reads). Thus, adjusting the threshold according to the per-locus allelic richness provides the most accurate estimate of absolute abundance in a mixed-DNA sample.

At the lowest threshold (0.001), contributor estimations of eDNA samples exhibited a positive bias across all densities, indicating the presence of false positive alleles in eDNA reads. However, the lack of cross-contamination in the negative control samples indicates the false positive alleles likely arose from artifacts introduced during PCR or sequencing, rather than cross-contamination between eDNA and tissue samples. The issue of positive bias in contributor estimations was more prevalent when population allele frequencies were specified using eDNA read frequencies. Nonetheless, patterns of bias and estimates of the number of genetic contributors in mesocosms were similar regardless of whether the input population allele frequencies were characterized using eDNA read frequencies or tissue-based allele frequencies (Figure 3.2). Thus, under controlled conditions, genotyped individuals may not be required to obtain reliable estimates of the absolute abundance of a species in eDNA samples.

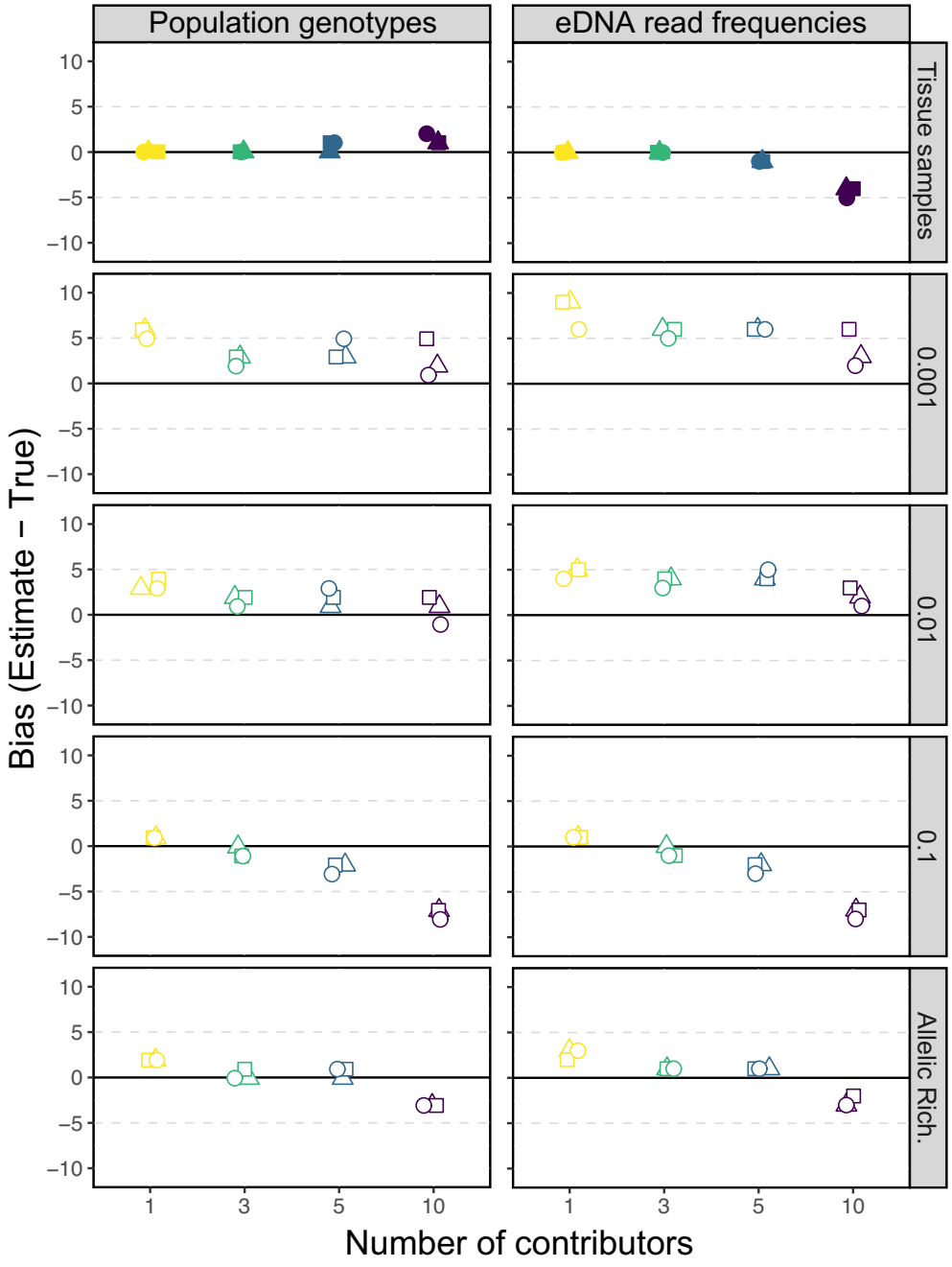

Figure 3.2. Bias of the contributor estimation using genotypes from Round Goby tissue samples (filled symbols) and eDNA samples (hollow symbols) across mesocosm treatments of Round Goby density (1, 3, 5, or 10 fish). The population allele frequencies for mixture estimation input were derived from 58 genotyped individual Round Gobies (left) or from

eDNA read frequencies combined across all mesocosms. Symbols represent treatment replicates and panels indicate fixed threshold frequencies below which sequence reads were removed (0.1, 0.01, 0.001) or a variable threshold based on per-locus allelic richness (Allelic Rich.).

In simulated mesocosm mixtures, the number of individuals could be reasonably estimated in mixtures up to 58 individuals, although bias associated with threshold values were apparent (Figure 3.3). The highest threshold (0.1) often exceeded the allele frequencies for all but the most common alleles, resulting in a negative bias in the contributor estimation. At this threshold, the maximum contributor estimation peaked at around 15 individuals, even at simulated densities > 50 individuals. On the other hand, filtering the eDNA reads according to lower thresholds (0.01, 0.001) appeared to overestimate the number of contributors across all densities. The variable threshold based on allelic richness showed lower variation in the estimated number of contributors and performed well for all but the largest numbers of contributors, where a negative bias occurred. Thus, while our 28-locus panel was able to resolve mixtures of up to 58 individuals, filtering decisions can have a large effect on the estimated number of contributors in eDNA samples.

In the field trial, the contributor estimation resulted in an estimated 5, 3, and 3 genetically distinct individuals captured by the three replicate eDNA samples when population-level allele frequencies *p* were estimated from the 73 genotyped individuals. However, because the contributor estimation calculations only consider alleles from the specified population-level allele frequencies, this is likely an underestimate as we did not recover several low-frequency alleles from the genotyped tissues in the eDNA samples. When

population-level allele frequencies were specified using the combined reads from the three replicate eDNA samples, an estimated 13, 7, and 6 genetically distinct individuals contributed to the mixture of DNA from each sample, respectively.

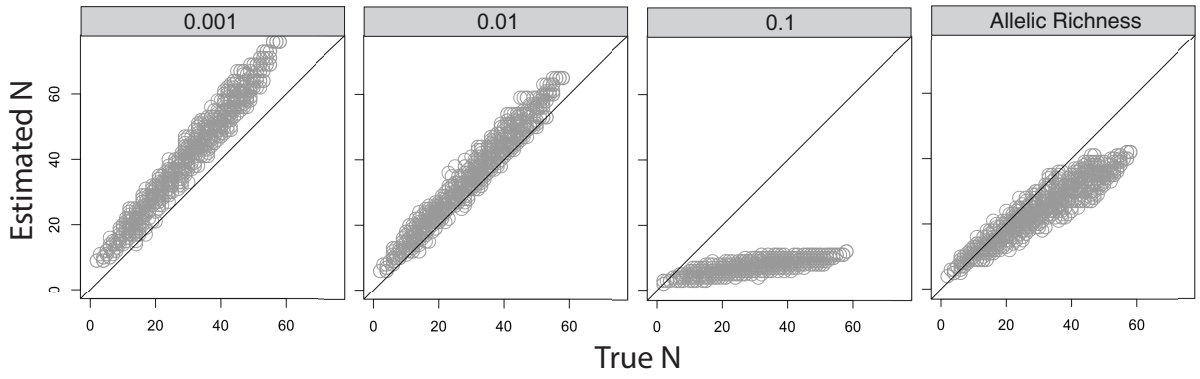

Figure 3.3. Estimated number of individuals contributing to simulated eDNA mixtures (range 2–58 individuals) using alleles from 1,000 simulated eDNA mixtures generated by constructing combinations of up to 12 mesocosms. Panels correspond to fixed threshold frequencies below which sequence reads were removed (0.001, 0.01, 0.1) or a variable threshold based on per-locus allelic richness. Diagonal lines represent a 1:1 relationship (i.e. zero bias for mixture contributor estimates).

Discussion

Estimating the genetic diversity and abundance of a species provides insights into a wide range of ecological and evolutionary processes and may have important implications for conservation management opportunities. While analysis of eDNA is a well-established approach for detecting species, it also holds potential to detect genetic diversity within species (Adams et al., 2019; Sigsgaard et al., 2020). With this study, we use eDNA and NGS methods

to detect intraspecific genetic diversity of an aquatic invasive species by recovering microsatellite allele frequencies that are similar to those derived from genotyped tissue samples in experimental mesocosms and in field-based eDNA collections. Using DNA mixture analyses, we estimated the number of genetic contributors of the target species within environmental samples, demonstrating the ability to use intraspecific genetic information to estimate the number of individuals captured in an eDNA sample. Although technical challenges regarding the parsing out of sequencing noise from low-abundance alleles in eDNA samples remain, this study experimentally validates the use of nuclear microsatellites to estimate population-level allele frequencies and absolute abundance of aquatic species using eDNA methods, a requisite step toward population-level inferences using nuclear eDNA.

To date, studies using eDNA approaches to characterize intraspecific genetic variation in aquatic species have been limited to single loci in the mitochondrial genome (Elbrecht et al. 2018; Parsons et al., 2018; Sigsgaard et al., 2017; Tsuji et al., 2019; Turon et al. 2019). The expansion of eDNA approaches to target multiallelic nuclear DNA markers could allow for the detection of robust higher-resolution population-level genetic information from water samples, as is common practice in contemporary tissue-based population genetics studies. In controlled mesocosms, we document microsatellite allele frequencies from eDNA closely resembled tissue-based allele frequencies across all mesocosms and on a per-mesocosm basis, although our approach exhibited decreased sensitivity in genetically distinguishing mesocosms from one another at high Round Goby densities (Figure 3.1C–D). Because we used Round Gobies derived from a single population source, this is to be expected. We also demonstrate reasonably accurate allele frequency estimates from eDNA samples collected in

natural conditions in a field trial, albeit with reduced detection of rare alleles in the population. Such eDNA-based estimates of population-level allele frequencies could potentially be used in population genetic inferences and demographic analyses using eDNA sampling methods. However, because eDNA samples contain a pool of DNA from many individuals, this approach is unable to determine multi-locus genotypes or assign genotypes to individuals, and methods designed to analyze population genetics using individual genotypes will need to be adapted into an eDNA framework. Theoretical and analytical frameworks for estimating population genetic parameters from pooled tissue samples of many individuals (Pool-seq) have already been developed (Boitard et al., 2013; Gautier et al., 2013; Hivert, Leblois, Petit, Gautier, & Vitalis, 2018), and similar frameworks may be useful for eDNA-based population genetics. As emphasized in Sigsgaard et al. (2020), however, such frameworks may need to account for additional potential sources of bias affecting the precision of population allele frequency estimates from eDNA, including variation in the number of individuals sequenced, unequal contributions of DNA from individuals, and variation from library preparation and sequencing.

Detecting intraspecific genetic variation in eDNA samples is also useful for estimating the number of genetically distinct individuals detected in a sample, which may be advantageous over approaches using DNA concentrations to predict species abundance or biomass. With the number of loci used in this study, the number of genetic contributors in simulated mixtures of up to 58 individuals could be resolved. While contributor estimations at the highest allele frequency threshold provided the most accurate estimates at low Round Goby densities in the mesocosm experiment, they were insufficient in resolving high numbers of Round Gobies, likely due to the removal of true low-frequency alleles below the threshold

limits. In contrast, low thresholds sufficiently resolved the number of contributors at high Round Goby densities but erroneously inflated the number of contributors at low densities due to the introduction of false alleles. We therefore recommend bioinformatic filtering based upon moderate thresholds or variable thresholds associated with locus-specific allelic richness in future applications of DNA mixture analysis. However, we also caution future studies to further investigate the possible impacts of false alleles and allelic dropout on contributor estimations, particularly in field-based settings where false alleles are more difficult to distinguish from true low-abundance alleles and detection of rare alleles may be low. Because low-frequency alleles provide strong information on the number of individuals present in a sample (Sethi et al. 2019), efforts to maximize the recovery of low-frequency alleles through optimization of field and laboratory protocols may be required to obtain accurate estimates of the number of individuals captured in eDNA samples. Additionally, applications of error-correction algorithms and denoising procedures may be required to aid in the detection and removal of erroneous sequences while retaining true low-frequency alleles (Elbrecht et al., 2018; Tsuji et al., 2019; Turon et al., 2019).

Future eDNA studies may consider the use of single nucleotide polymorphisms (SNPs) as a target nuclear marker, as they are an abundant and widespread form of variation throughout the genome of most species (Morin, Luikart, & Wayne, 2004). However, because the inferential power of the DNA mixture model is limited by the number of recovered alleles, much larger marker panels of bi-allelic SNPs will be needed to resolve eDNA mixtures into the number of genetic contributors, particularly as the number of contributors grows (Sethi et al., 2019). Rather than targeting single SNPs, a potential solution may be to target several SNPs occurring in the same genomic region that can be jointly genotyped (Kidd et al., 2013).

Such multi-allelic "microhaplotype" markers have high per-locus information content in a small length of DNA and may reduce the potential for analysis errors that arise when targeting microsatellites including PCR stutter and allelic dropout.

Although our approach demonstrates promise for future applications of non-invasive population genetic sampling using nuclear eDNA, the controlled settings of our mesocosm experiments and limited spatial and temporal scale of the field trial may not reflect the complexity of *in situ* conditions. Thus, several obstacles may need to be addressed before this approach can be broadly applied in field-based settings. For instance, although Round Gobies may exhibit localized hotspots of high density, the average density of Round Gobies in occupied habitats of Cayuga Lake (1.82 fish/m^2) is lower than in our mesocosm experiments (Andres et al., 2020), and read depths we observed in mesocosm eDNA samples may not be achievable in field settings. Indeed, even with targeted eDNA sampling in areas of high expected Round Goby densities, read depths in eDNA samples from the field trial averaged 4,305 reads per sample, which is much lower than reported in other eDNA studies using targeted field sampling and markers in the mitochondrial genome (e.g., average 263,111 reads per sample at sites where whale sharks were reported, Sigsgaard et al., 2016; average 237,434.5 reads per sample taken from harbour porpoise fluke prints, Parsons et al., 2018). To ensure genetic data obtained from eDNA samples sufficiently reflects the genetic diversity of the population of interest when targeting loci in the nuclear genome, efforts to evaluate the limit of detection and optimize field and laboratory strategies to achieve sufficient eDNA copy numbers may be required (Adams et al., 2019; Sigsgaard et al., 2020).

Mesocosm conditions also lacked the biophysical complexity inherent in natural systems, where many other particles and organisms are present and contributing to eDNA

samples (Barnes & Turner, 2015). PCR inhibition from non-target particles may restrict accurate molecular identification of alleles, particularly when coupled with low eDNA concentrations of target species DNA (Hunter, Ferrante, Meigs-Friend, & Ulmer, 2019). Importantly, if closely related non-target species are found in the sampled habitats, primer specificity must be thoroughly tested to ensure DNA from co-occurring non-target species is not amplified. While no congeners of the Round Goby are found in North America, the freshwater Tubenose Goby (*Proterorhinus semilunaris*, formerly known as *P. marmoratus*; Stepien & Tumeo 2006) is found throughout the Great Lakes. Although we tested primer specificity *in silico* using DNA databases, *in vitro* testing using tissue-derived DNA from non-target species may also be required if reference sequence data is lacking for closely related co-occurring species.

With proper validation and appropriate analytical frameworks, eDNA-based population genetics has the potential to enhance the use of eDNA methods in conservation and management of species. For example, preventing the spread and minimizing the undesirable impacts of invasive species will require effective monitoring of non-native populations, including evaluating population-level genetic variation and population size at the sites of initial colonization (Le Roux & Wieczorek, 2009). With further development, this method might someday inform management strategies at early stages in the invasion process when eradication efforts are most likely to be successful in preventing proliferation and future spread (Leung et al., 2002; Lodge et al., 2016). This approach may also be beneficial for monitoring species where small population sizes, expansive or complex habitats, elusive behavior, or a desire to minimize invasive sampling can prevent effective population assessments. For instance, Parsons et al. (2018) used eDNA approaches to generate

mitochondrial sequence data in a highly elusive marine mammal where physical tissue-based sampling presents logistical challenges and limits the detection of population genetic structure. The high sensitivity of eDNA methods and relative ease of sample collection therefore present a noninvasive and potentially cost-effective opportunity to study the population genetics of aquatic organisms for which traditional sampling is difficult or impossible.

As with other eDNA methods such as DNA metabarcoding, the approach developed here is likely to complement, rather than replace, existing methods of evaluating intraspecific diversity in population genetics studies (Yoccoz, 2012). Indeed, developing species-specific panels of microsatellite DNA markers requires sufficient DNA sequence data for the species of interest, and optimization of multiplex PCR requires testing on tissue-derived DNA samples. Estimating the number of contributors to eDNA samples also requires an assessment of population allele frequencies, an estimate that may be derived from tissue-based genotyping of the population of interest. However, we demonstrate that under controlled experimental conditions, population allele frequencies from eDNA read frequencies are highly correlated with allele frequencies from genotyped individuals and contributor estimations are similar regardless of where the population allele frequencies are derived. Estimating the number of contributors in eDNA samples may therefore be feasible even in the absence of population-level sequence information from tissue samples.

As the time and costs associated with obtaining and analyzing molecular data continue to decline, eDNA methodologies may become an increasingly effective approach for detecting and quantifying the presence of invasive, rare, or threatened species. Moreover, with the recent expansion of eDNA approaches into studies of intraspecific diversity, the scope of

eDNA applications has broadened to population-level inferences. With this study, we demonstrate the advancement of eDNA approaches to encompass genetic markers in the nuclear genome, with implications for future studies of population genetics using next-generation sequencing of environmental samples. By incorporating DNA mixture analyses into a nuclear genome-based eDNA framework, we estimate the number of unique contributors to eDNA samples, providing the first steps to a potential alternative to correlation-based estimates of species abundance using DNA concentration. Furthermore, we demonstrate the ability to obtain population-level genetic information from nuclear eDNA, supporting the potential for future assessments of population genetics from environmental samples. Provided further validation and optimization in field-based settings, such an advancement could transform the ways in which we obtain population-level genetic information on species of conservation or management concern.

Acknowledgements

This study was supported with funding provided by the National Science Foundation (NSF) Coastal SEES program (grant number 1748389) and the Department of Defense (DoD) Strategic Environmental Research and Development Program (SERDP) (grant number RC19-1004). We thank Amanda Wong and Amelia Weiss for assistance with collecting specimens. We also thank Wes Larson and Timothy Lambert for discussions regarding the application and analysis of contributor estimations using eDNA samples. We are grateful to three anonymous reviewers for their valuable comments on the manuscript. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G. J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool?. *Genes*, *10*(3), 192.

Andres, K. J., Sethi, S. A., Duskey, E., Lepak, J. M., Rice, A. N., Estabrook, B. J., ... Perkins, K. (2020). Seasonal habitat use indicates that depth may mediate the potential for invasive Round Goby impacts in inland lakes. *Freshwater Biology*. *65*(8), 1337-1347.

Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics, 17*(1), 1-17. doi:10.1007/s10592-015-0775-4.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, 13, 729–744.

Bohmann, K., Evans, A., Gilbert, M. T., Carvalho, G. R., Creer, S., Knapp, M., ... de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution, 29*(6), 358-367. doi:10.1016/j.tree.2014.04.003

Boitard, S., Kofler, R., Françoise, P., Robelin, D., Schlötterer, C., & Futschik, A. (2013). Pool-hmm: A Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Molecular Ecology Resources*, 13, 337–340.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30* 2114-2120.

Bylemans, J., Furlan, E. M., Hardy, C. M., McGuffie, P., Lintermans, M., & Gleeson, D. M. (2017). An environmental DNA-based method for monitoring spawning activity: a

case study, using the endangered Macquarie perch (*Macquaria australasica*). *Methods in Ecology and Evolution, 8*(5), 646-655.

Charlebois, P. M., Marsden, J. E., Goettel, R. G., Wolfe, R. K., Jude, D. J., & Rudnika, S. (1997). *The Round Goby,* Neogobius melanostomus *(Pallas), a review of European and North American Literature* (Vol. INHS Special Publication No. 20): Illinois-Indiana Sea Grant Program and Illinois Natural History Survey.

Curran, J. M., Triggs, C. M., Buckleton, J., & Weir, B. (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Science, 44*(5), 987-995.

D'Aloia, C. C., Bogdanowicz, S. M., Harrison, R. G., & Buston, P. M. (2017). Cryptic genetic diversity and spatial patterns of admixture within Belizean marine reserves. *Conservation Genetics*, *18*(1), 211-223.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... De Vere, N. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution, 8*(12), 1888-1898.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology, 49*(4), 953-959.

Elbrecht, V., Vamos, E.E., Steinke, D. and Leese, F., 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, 6, p.e4644.

Faircloth, B. C. (2008). MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources, 8*(1), 92-94.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters, 4*(4), 423-425. doi:10.1098/rsbl.2008.0118

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., … Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, 22, 3766–3779.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., ... Cornman, R. S. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307.

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., ... Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology, 25*(13), 3101-3119.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, 210, 315–330.

Hunter, M. E., Ferrante, J. A., Meigs-Friend, G., & Ulmer, A. (2019). Improving eDNA yield and inhibitor reduction through increased water volumes and multi-filter isolation techniques. *Scientific Reports*, *9*(1), 1-9.

Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited

symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525-1534.

Iversen, L. L., Kielgast, J., & Sand-Jensen, K. (2015). Monitoring of animal abundance by environmental DNA — An increasingly obscure perspective: A reply to Klymus et al., 2015. *Biological Conservation, 192*, 479-480. doi:https://doi.org/10.1016/j.biocon.2015.09.024

Janssen, J., & Jude, D. J. (2001). Recruitment failure of mottled sculpin *Cottus bairdi* in Calumet Harbor, southern Lake Michigan, induced by the newly introduced round goby *Neogobius melanostomus*. *Journal of Great Lakes Research, 27*(3), 319-328. doi:10.1016/s0380-1330(01)70647-8

Jerde, C. L., Mahon, A. R., Chadderton, W. L., & Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters, 4*(2), 150-157.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24, 1403-1405. doi: 10.1093/bioinformatics/btn129.

Jude, D. J., Reider, R. H., & Smith, G. R. (1992). Establishment of Gobiidae in the Great Lakes Basin. *Canadian Journal of Fisheries and Aquatic Sciences, 49*(2), 416-421. doi:10.1139/f92-047

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS One, 9*(1), e86175.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific reports*, *9*(1), 1-14.

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., & Ihuegbu, N. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, *4*(1), e123-e124.

Klymus, K. E., Richter, C. A., Chapman, D. C., & Paukert, C. (2015). Quantification of eDNA shedding rates from invasive bighead carp *Hypophthalmichthys nobilis* and silver carp *Hypophthalmichthys molitrix*. *Biological Conservation, 183*, 77-84.

Kornis, M. S., Mercado-Silva, N., & Vander Zanden, M. J. (2012). Twenty years of invasion: a review of Round Goby *Neogobius melanostomus* biology, spread and ecological implications. *Journal of Fish Biology, 80*(2), 235-285. doi:10.1111/j.1095-8649.2011.03157.x

Krakowiak, P. J., & Pennuto, C. M. (2008). Fish and macroinvertebrate communities in tributary streams of eastern Lake Erie with and without Round Gobies (*Neogobius melanostomus*, Pallas 1814). *Journal of Great Lakes Research, 34*(4), 675-690. doi:doi.org/10.3394/0380-1330-34.4.675

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: variability among capture methods and environmental conditions. *Molecular Ecology Resources, 16*(6), 1401-1414.

Le Roux, J., & Wieczorek, A. (2009). Molecular systematics and population genetics of biological invasions: towards a better understanding of invasive species management. *Annals of Applied Biology, 154*(1), 1-17.

Leung, B., Lodge, D. M., Finnoff, D., Shogren, J. F., Lewis, M. A., & Lamberti, G. (2002). An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive

species. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 269*(1508), 2407-2413.

Lodge, D. M., Simonin, P. W., Burgiel, S. W., Keller, R. P., Bossenbroek, J. M., Jerde, C. L., … Zhang, H. (2016). Risk analysis and bioeconomics of invasive species to inform policy and management. *Annual Review of Environment and Resources, 41*, 453-488.

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., … Pfrender, M. E. (2012). Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology, 21*(11), 2555-2558.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS One, 9*(12), e114639.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp Cyprinus carpio. *Molecular ecology resources*, *17*(2), 324-333.

Morin, P. A., Luikart, G., & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in ecology & evolution*, *19*(4), 208-216.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics, 26*(3), 419-420.

Parsons, K. M., Everett, M., Dahlheim, M., & Park, L. (2018). Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. *Royal Society open science, 5*(8), 180537.

Piggott, M. P. (2016). Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish. *Ecology and evolution*, *6*(9), 2739-2750.

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., & Waits, L. P. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences, 70*(8), 1123-1130.

R Core Team. (2016). R: A language and environment for statistical computing.

Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. In: *Bioinformatics methods and protocols* (pp. 365-386): Springer.

Rubinoff, D., Cameron, S., & Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of heredity*, *97*(6), 581-594.

Schaeffer, J. S., Bowen, A., Thomas, M., French III, J. R., & Curtis, G. L. (2005). Invasion history, proliferation, and offshore diet of the Round Goby *Neogobius melanostomus* in western Lake Huron, USA. *Journal of Great Lakes Research, 31*(4), 414-425.

Sethi, S. A., Larson, W., Turnquist, K., & Isermann, D. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution, 10*(1), 109-119.

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., … Willerslev, E. (2017). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution, 1*(1), 0004.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245-262.

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S., … Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution, 8*(5), 635-645.

Stepien, C. A. & Tumeo, M. A. (2006). Invasion genetics of Ponto-Caspian gobies in the Great Lakes: a 'cryptic'species, absence of founder effects, and comparative risk analysis. *Biological Invasions*, *8*(1), 61-78.

Stewart, K., Ma, H., Zheng, J., & Zhao, J. (2017). Using environmental DNA to assess population-wide spatiotemporal reserve use. *Conservation Biology, 31*(5), 1173-1182.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology, 21*(8), 1789-1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology, 21*(8), 2045-2050.

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., & Kawabata, Z. i. (2012). Estimation of fish biomass using environmental DNA. *PLoS One, 7*(4), e35868.

Teske, P. R., Golla, T. R., Sandoval-Castillo, J., Emami-Khoyi, A., van der Lingen, C. D., von der Heyden, S., … Beheregaray, L. B. (2018). Mitochondrial DNA is unsuitable to test for isolation by distance. *Scientific Reports, 8*(1), 8448.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One, 7*(8), e41732.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation, 183*, 4-18. doi:10.1016/j.biocon.2014.11.019

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, *2*(1), 42-52.

Turon, X., Antich, A., Palacín, C., Præbel, K. and Wangensteen, O.S., 2020. From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications*, 30(2), p.e02036.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Vander Zanden, M. J., Hansen, G. J., Higgins, S. N., & Kornis, M. S. (2010). A pound of prevention, plus a pound of cure: early detection and eradication of invasive species in the Laurentian Great Lakes. *Journal of Great Lakes Research, 36*(1), 199-205.

Vélez-Espino, L. A., Koops, M. A., & Balshine, S. (2010). Invasion dynamics of Round Goby (*Neogobius melanostomus*) in Hamilton Harbour, Lake Ontario. *Biological Invasions, 12*(11), 3861-3875. doi:10.1007/s10530-010-9777-9

Weir, B. S., Triggs, C., Starling, L., Stowell, L., Walsh, K., & Buckleton, J. (1997). Interpreting DNA mixtures. *Journal of Forensic Science, 42*(2), 213-222.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M. B., Brand, T. B., … Dahl-Jensen, D. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science, 317*(5834), 111-114.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., … Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science, 300*(5620), 791-795.

Yoccoz, N. G. (2012). The future of environmental DNA in ecology. *Molecular Ecology, 21*(8), 2031-2038.

CHAPTER 4

ENVIRONMENTAL DNA DETECTS POPULATION GENETIC DIVERSITY AND STRUCTURE WITH MICROSATELLITE MARKERS

Abstract

Environmental DNA (eDNA) has been established as a noninvasive and efficient approach to sample genetic material from aquatic environments. Although most commonly used to determine species presence and measure biodiversity, eDNA samples may also hold great potential to obtain population-level genetic information from water samples. In this study, we sequenced a panel of multiallelic microsatellite markers from eDNA and tissue samples to uncover intraspecific diversity in the nuclear genome of the Round Goby (*Neogobius melanostomus*) at 13 locations across their invaded range. We tested for patterns of population genetic diversity and pairwise genetic distance using eDNA-based and individual genotype-based estimates of allele frequencies and assessed the similarity in population genetic inferences calculated using the two approaches. eDNA approaches detected alleles at similar frequencies to genotyped tissues, with correlations of up to $r = 0.85$ between the relative frequencies of eDNA sequences and the relative frequencies of alleles in sampled individuals. Genetic variability within and between populations was also detected from eDNA in patterns that were consistent with individual tissue-based estimates of genetic diversity and differentiation, with the strongest differentiation occurring along the East-West gradient in an isolation by distance pattern. Our study demonstrates the potential for eDNA-based approaches to characterize key population parameters required to effectively monitor, manage, or sustain aquatic species.

Introduction

The analysis of environmental DNA (eDNA) has revolutionized our ability to noninvasively study species composition and measure biodiversity (Deiner et al., 2017). By capturing DNA fragments present in environmental samples (i.e., soil, water, and air), eDNA approaches can identify the source organisms that contribute genetic information to a mixed eDNA sample. Although eDNA was initially used for the detection of single species, developments in eDNA metabarcoding now allow for hundreds of species to be detected simultaneously from a single sample, with wide-ranging uses including biodiversity assessments and biomonitoring (Thomsen & Willerslev, 2015; Valentini et al., 2016). Furthermore, because eDNA approaches often have higher sensitivity than conventional capture-based methods for detecting species (Evans, Shirey, Wieringa, Mahon, & Lamberti, 2017; Fediajevaite, Priestley, Arnold, & Savolainen, 2021), they represent a rapid and cost-effective approach for detecting rare, elusive, and invasive species or for characterizing entire biological communities (Sepulveda et al., 2019; Thomsen et al., 2012).

While eDNA is typically used for detecting taxonomic diversity at the species level, recent studies show promise for detecting within-species genetic variation as well (reviewed in Adams et al., 2019; Sigsgaard et al., 2020). For instance, Uchii, Doi, and Minamoto (2016) used a cycling probe qPCR assay targeting a single SNP on the mitochondrial d-loop to quantify the proportion of native and non-native genotypes of Common Carp (*Cyprinus carpio*). Intraspecific genetic diversity has also been revealed through eDNA metabarcoding by sequencing short variable regions of the mitochondrial genome, with haplotype variation characterized with eDNA approaches reflecting haplotype variation in conventional tissue-based genetic approaches (Parsons, Everett, Dahlheim, & Park, 2018; Sigsgaard et al., 2016).

Much longer mitochondrial DNA (mtDNA) fragments may also be recovered in eDNA samples, with high coverage sequence data spanning entire mitogenomes possible using target capture probes (Jensen et al., 2021). Such research challenges the paradigm that only short degraded fragments are present in eDNA samples and demonstrates the potential for longer and more variable DNA markers to be used for population-level assessments using eDNA.

While mitochondrial haplotypes have been detected from eDNA in several studies, analyses of nuclear DNA (nuDNA) are relatively rare, likely due to the precedent of using mtDNA markers in eDNA metabarcoding and the expected lower concentration of nuDNA in environmental samples (Sigsgaard et al., 2020). However, because the mitochondrial genome typically does not recombine, mtDNA markers represent a single independent locus and, in many cases, may not contain sufficiently detailed genetic information to characterize population genetic patterns (Ballard & Whitlock, 2004). To obtain higher-resolution population genetic information from eDNA samples, sequence variation at multiple independent nuclear genetic markers should be assessed. Although Andres, Sethi, Lodge, and Andrés (2021) recently showed that nuclear microsatellites can be detected in experimental mesocosms and a field sample, it remains to be seen if field collections of eDNA can provide population genetic information using nuclear genetic markers.

To investigate the potential for eDNA approaches to uncover intraspecific genetic variation in field settings, we conducted an eDNA-based and tissue-based population genetic assessment of the Round Goby (*Neogobius melanostomus*) in sites across the Great Lakes region. Round Goby is an invasive fish species that was first discovered in the St. Clair River in 1990, probably having been introduced via ballast water discharge from shipping vessels coming from the Ponto-Caspian region of eastern Europe, where Round Gobies are native

(Jude, Reider, & Smith, 1992). The invaded range of Round Gobies has since expanded to include all five Great Lake and many of their surrounding tributaries, rivers, and canals, with secondary spread occurring through a combination of natural dispersal and ongoing ballast exchange within the Great Lakes region (Bronnenhuber, Dufour, Higgs, & Heath, 2011; George, Baldigo, Rees, Bartron, & Winterhalter, 2021; Ricciardi & MacIsaac, 2000). Round Gobies have variable ecological and economic impacts in invaded ecosystems, as they prey upon and compete with native species (Balshine, Verma, Chant, & Theysmeyer, 2005; Janssen & Jude, 2001), consume invasive dreissenid mussels (Johnson, Bunnell, & Knight, 2005), and serve as an important prey item for several fisheries species (Crane & Einhouse, 2016; Taraborelli, Fox, Johnson, & Schaner, 2010). Understanding the invasion history and identifying the factors governing their ongoing spread are important for predicting the potential distribution and impacts of Round Gobies, and such insights may be revealed with the use of molecular genetic data.

Previous tissue-based population genetic analysis of mitochondrial and nuclear markers has revealed high levels of genetic diversity and structuring among populations in the Round Goby invasive range, with no indication of founder effects or a population bottleneck during introduction, suggesting high propagule pressure from multiple source populations (Brown & Stepien, 2009; Stepien & Tumeo, 2006). Sites within and among the Great Lakes experience restricted post-establishment gene flow, with limited natural dispersal paired with occasional long-distance dispersal events likely mediated by vessel transport among Great Lakes sites (LaRue, Ruetz, Stacey, & Thum, 2011). More recent dispersal into adjacent tributaries occurs via a stratified dispersal strategy, with a combination of short-distance contiguous dispersal and long-distance migrants (Bronnenhuber et al., 2011). Because the

Round Goby invasion exhibits high levels of diversity and genetic structure, it provides an ideal system for developing eDNA-based population genetics tools, as we can test the potential for eDNA approaches to detect hypothesized genetic patterns including increased levels of genetic diversity found at the invasion origin (i.e., leading edge hypothesis), significant population structure, and isolation by distance.

With this research, we use eDNA analysis to uncover population genetic patterns of Round Gobies within and across sites spanning their invaded range. We ask: (1) Can nuclear microsatellite markers for Round Gobies be amplified from environmental samples? (2) Do allele frequencies from eDNA reflect allele frequencies from tissue-based genetic analysis? (3) Do population genetic patterns inferred from eDNA samples align with genetic patterns established with the analysis of Round Goby tissues? Because the ability to detect and amplify single-copy nuDNA markers from environmental samples has not been well studied, we also investigated the relative concentrations of Round Goby mtDNA and nuDNA in eDNA samples. While the primary goal of this study is to investigate the ability to detect and analyze nuclear microsatellites using eDNA approaches, our study design includes sites in the Great Lakes that were sampled in Brown and Stepien (2009) as well as several new sites at the Round Goby invasion front in the New York State Canal System. We therefore provide an updated examination of the population genetic patterns of this invader that may help to identify patterns of secondary spread and contribute to understanding the ecological success of invasive species in new habitats.

Materials and methods

Sample collection and DNA extraction

Water samples and Round Gobies were collected in July–September 2019 at each of 13 locations across the invasive range of Round Gobies (Table S4.1). Sampled sites encompass locations across four of the five Laurentian Great Lakes (Michigan, Huron, Erie, and Ontario), a site near the first documented Round Goby introduction in Lake St. Clair, and sites along the more recent eastward invasion front in the Erie Canal and Finger Lakes region. Microsatellite allele frequencies from eDNA and tissue samples collected from Cayuga Lake, NY were analyzed in Andres et al. (2021) and are included in this study. We collected eDNA samples without accompanying Round Goby tissues at one site on Seneca Lake, NY where Round Gobies are not yet known to occur (Freedman et al., 2022).

At each site, triplicate 2-L surface water samples were collected from shore in sterile collection bags (Whirl-Pak) and filtered through Whatman cellulose nitrate membrane filters (47 mm diameter; 1 μm pore size) using a hand-operated vacuum pump and filter funnel (Pall Corporation). If at any time the filters became clogged or developed cracks due to pressure, filtration was halted, the filter was replaced, and the remaining water was filtered through a new membrane. We filtered 2 L of distilled water at each site to serve as negative controls. All membrane filters were preserved in 700 μL of Longmire's buffer (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5% SDS) and kept at room temperature for up to one week before storage at -20°C until DNA extraction. Prior to sampling, all reusable eDNA sampling and filtration equipment was cleaned with 50% household bleach, rinsed with DI water, and treated under UV light for 30 minutes. Immediately following eDNA sample collection, up to 30 Round Goby individuals were collected at each site (N = 299; Table S4.1) via beach seining. Fish were humanely euthanized with MS-222. Fin tissue was sampled from the caudal fin of each fish and stored in 95% EtOH until DNA extraction.

All pre-PCR laboratory work for eDNA samples was carried out in a dedicated PCR-free laboratory where all surfaces and reusable laboratory equipment (e.g., pipettes, tube racks) are bleached after each use and UV-irradiated daily. This laboratory is housed in a separate facility from the laboratory used to handle post-PCR eDNA samples. All eDNA laboratory work was carried out in a separate facility from the laboratory used to process Round Goby tissues.

For Round Goby tissues, we extracted genomic DNA using the HotSHOT protocol (Truett et al., 2000), which included incubating goby fin clips with 25 mM NaOH at 95°C for 30 min followed by neutralization with 40 mM Tris-HCl. The DNeasy Blood and Tissue extraction kit (Qiagen Inc.) was used to extract eDNA with a modified protocol that includes centrifugation of the preservation buffer, increased volumes of lysis and digestion reagents, and extended incubation periods to accommodate eDNA filters (Spens et al., 2017). A negative control was added at this stage by performing the same DNA extraction steps on 700 µL of molecular grade deionized water in place of Longmire's buffer. If more than one filter was used for a sample, each filter was extracted separately and the elution was pooled into a single sample. The final elution volume for all eDNA samples was 100 µL. Extracted DNA from eDNA samples and Round Goby tissues was stored at -20°C until amplification.

qPCR assays

To test the relative quantities and detectability of mtDNA and nuDNA markers in eDNA samples, we measured the quantity of each type of DNA (copies/L of sampled water) in each sample using species-specific primer sets and fluorescent Taqman probes in quantitative PCR (qPCR) assays. The mtDNA marker, developed for previous research on

eDNA-based detection of Round Gobies in their invasive range (Nathan, Jerde, Budny, & Mahon, 2015; Nevers et al., 2018), targets a 147 bp region of the Round Goby cytochrome oxidase I (COI) gene (Table S4.2). Standard curves were generated from 10-fold serial dilutions (10–10^5 copies/µL) of a gBlock Gene Fragment (Integrated DNA Technologies). We designed the gBlock fragment based on the consensus sequence of aligned Round Goby COI sequences from GenBank (https://www.ncbi.nlm.nih.gov/genbank/), with the synthesized 191 bp DNA fragment overlapping the target sequences by at least 16 bp on either side. To identify possible sample cross-contamination by the gBlock fragment, position 61 of the synthesized DNA contains adenine in place of guanine, resulting in a stop codon that would not occur in DNA sequences derived from Round Gobies.

To quantify nuclear eDNA, we developed a marker for this study targeting a short, conserved region (60 bp) of the *Nmel505* microsatellite locus from Andres et al. (2021). After aligning all alleles of this locus from genotyped Round Goby tissues in Geneious Prime (2020.0.5), we designed qPCR primers and a Taqman probe using Primer Express (v3.0.1). As with the mtDNA assay, we designed a gBlock fragment (125 bp) that encompassed the entire target sequence. A 10-fold serial dilution (10–10^5 copies/µL) of the synthesized gBlock was used to generate a standard curve, allowing for the quantification of nuDNA copies/µL in eDNA samples.

We ran qPCR assays on the QuantStudio 7 Pro at the Cornell Institute of Biotechnology's Genomics Facility with three technical replicates of each eDNA sample, standard dilutions, and no-template negative controls in each run. Reactions were prepared in 20 µL volumes containing 10 µL TaqMan Environmental Master Mix 2.0 (ThermoFisher Scientific), 1.8 µL each 10-µM primer (forward and reverse), 0.05 µL of 100-µM Taqman

probe, 4.35 µL of PCR-grade water, and 2 µL template DNA. Thermocycler conditions included an initial denaturation step at 95°C for 10 min followed by 45 cycles of 95°C for 15 s and either 60°C (mtDNA) or 58°C (nuDNA) for 1 min. We used the Design and Analysis software (v2.6.0) to determine amplification efficiency and R^2 for each assay. The average DNA concentration was calculated for each sample as the average of the three technical replicates and expressed as copy number/µL of sample. Any replicate with no measurable amplification was treated as having zero DNA copies.

Statistical analysis for qPCR results and the remainder of this study was performed using R v4.1.2 (R Core Team, 2021) unless otherwise indicated. We tested the difference between mtDNA and nuDNA concentration in eDNA samples using a linear mixed model with the 'lmer' function in the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015), with sample and site specified as nested random effects. The correlation between the amount of mtDNA and nuDNA was tested using Pearson's correlation test and a linear regression was fitted to the log-transformed data.

Microsatellite PCR, library preparation, and sequencing

To investigate genetic diversity in eDNA samples, we amplified a panel of 28 microsatellite markers for Round Gobies developed in Andres et al. (2021). Each primer set targets a variable-length region of the nuclear genome, with microsatellite alleles ranging in size from 228 to 435 bp (Table S4.3). Primer pairs were grouped into multiplexes and eDNA samples were amplified with each multiplex assay in triplicate to account for PCR stochasticity. Each reaction contained: 2 µL of each primer in the multiplex in equimolar concentrations (2 µM), 10 µL of Multiplex PCR Master Mix (Qiagen Inc.), 2 µL bovine

serum albumin (BSA; 4 μg/μL), 3 μL PCR-grade water, and 3 μL of eDNA sample. A negative PCR control contained 3 μL molecular H_2O in place of DNA. The program for multiplex PCR is: initial denaturation at 95°C for 15 min followed by 40 cycles of 94°C for 30 s, 58°C for 3 min, and 72°C for 60 s. PCR products were visualized with electrophoresis on a 1% agarose gel stained with ethidium bromide. Equal volumes (5 μL) of each technical replicate and multiplex were combined for each sample and uniquely indexed in a second-stage PCR using Illumina Nextera XT tags. This amplification step included 1.5 μL PCR product, 4 μL of One*Taq* 5x buffer (New England Biolabs, Inc.), 0.4 μL of 10 mM dNTPs, 0.1 μL of One*Taq* DNA polymerase, 12.4 μL of molecular H_2O, and 0.8 μL of each Index 1 (i7) and Index 2 (i5) adapters. The index PCR thermocycling program included six cycles of 95°C for 30 s, 62°C for 60 s, and 68°C for 60 s, with a final extension step of 68°C for 10 min. We pooled the final amplicons of each sample and negative control (field, extraction, and PCR) in equal volumes and purified the eDNA library with Agencourt AMPure XP beads (reaction ratio AMPure beads 0.7: PCR product 1; Beckman Coulter Genomics). We assessed library concentration with a Qubit 2.0 fluorometer (Thermo Fisher Scientific), diluted to 18 nM, and sequenced at Cornell University's Institute of Biotechnology Genomics Facility on the Illumina MiSeq platform with the MiSeq v2 500 bp kit (paired-end 2 × 250 bp). Due to low read counts in eDNA samples, we re-sequenced the eDNA library in two additional MiSeq runs.

Multiplex PCR amplification of Round Goby tissues followed the same protocol as eDNA samples but with reduced reaction volumes (1 μL of 2 μM primer pairs, 5 μl of Qiagen Multiplex PCR Master Mix, 3 μL PCR-grade water, and 1 μL of Round Goby DNA) and reduced PCR cycles (initial denaturation at 95°C for 15 min and 35 cycles of 94°C for 30 s,

58°C for 90 s, and 72°C for 60 s). PCR products from each multiplex were pooled (5 μL) and each sample was barcoded using the same index PCR protocol described above. The tissue library was pooled, purified, and diluted as above and sequenced with a single MiSeq v2 500 bp kit (paired-end 2 × 250 bp).

Bioinformatic analysis

We processed a total of 81,822,102 demultiplexed sequence reads from eDNA samples and 30,255,948 reads from Round Goby tissue samples. Sequencing adapters were trimmed with Trimmomatic v0.39 (Bolger, Lohse, & Usadel, 2014) and sequence quality was assessed with FastQC version 0.11.8 (Andrews, 2010). We analyzed all eDNA and tissue sequences using a custom script (https://bitbucket.org/cornell_bioinformatics/amplicon/src/master/amplicon_dada2.py) that involves a modification to the DADA2 pipeline (Callahan et al., 2016) to allow for the processing of multiple loci. The script splits sequences based on primers, removes primers from forward and reverse reads, and discards sequences that are shorter than 100 bp or that have more than 2 expected errors (EE, calculated based on Phred scores). The DADA2 filtering algorithm was then used to identify and discard sequence errors from each sequence file. Denoised reads were merged with an overlap of at least 20 bp and 1 allowable mismatch in the overlap region. To filter out sequencing artifacts, all sequence variants at each locus were aligned to the most common allele from Round Goby tissues using the Smith-Waterman alignment algorithm on the first and last 25 bp of the query and reference sequences. The resulting dataset contained read counts of alleles at each locus for each tissue and eDNA sample.

Round Goby tissue samples were genotyped at each locus as described in Andres et al. (2021). At any given locus, individuals were designated as heterozygous if the read ratio between the two most common alleles was at least 0.2:0.8; otherwise, the individual was considered homozygous. Individuals with fewer than 10 reads at a locus were specified as having missing data at that locus. We removed one locus with > 50% missing data across all individuals (*Nmel*660; Table S4.3). The remaining 27 loci exhibited an average allelic richness of 14.4 alleles (range 3–29) with a total of 389 alleles across all loci. Reads for eDNA samples were combined across sequencing runs. eDNA read frequencies were then calculated per sample and per site by combining normalized reads from the three eDNA replicates.

Data analysis

To investigate patterns of population genetic diversity and structure in the invasive range of Round Gobies, we first conducted population genetic analyses using high-quality data from individual tissues. Allele frequencies for all genotyped individuals and for each population were calculated in `adegenet` (Jombart, 2008) and allelic richness, expected and observed heterozygosity, and the number of private alleles per population was calculated. To test whether time since invasion predicts genetic diversity, we determined the minimum time since invasion using the earliest published record of Round Goby presence within a 10-km radius of each sampled population (Freedman et al., 2022) and assessed the fit with linear models for the effect minimum time since invasion on allelic richness and number of private alleles. We calculated pairwise estimates of genetic diversity among sampled sites (F_{ST}; Weir & Cockerham, 1984) and assessed significance using 95% confidence intervals calculated

using a bootstrap procedure with 1000 bootstraps in `hierfstat` (Goudet, 2005). An assessment of the correlation between pairwise genetic distances and Euclidean geographic distances (measured in `geodist` Padgham & Sumner, 2021) was conducted using a Mantel test with 999 permutations in `ade4` (Dray & Dufour, 2007). Spatial patterns of genetic structure were visualized using Discriminant Analysis of Principal Components (DAPC) in `adegenet` (Jombart, Devillard, & Balloux, 2010) and further inspected using Bayesian assignment tests in STRUCTURE v2.3 (Pritchard, Stephens, & Donnelly, 2000) using an admixture model with correlated allele frequencies among groups. For each value of K (i.e., potential number of genetic clusters) from 1–13, we ran fifteen independent runs of 10^6 iterations of a Markov chain Monte Carlo simulation following a burn-in period of 10^5 iterations. We determined the likely number of genetic clusters according to the Delta-K statistic (Evanno, Regnaut, & Goudet, 2005). Hierarchical analysis of molecular variance (AMOVA) was performed in `poppr` (Kamvar, Tabima, & Grünwald, 2014) to determine the proportion of genetic variation that could be attributed to the genetic clusters identified in STRUCTURE and sampling site.

We then assessed the ability of eDNA to detect these patterns of genetic diversity and structure by evaluating the similarity between eDNA-based and tissue-based estimates of allele frequencies and population genetic parameters. Because eDNA contains genetic information from potentially multiple individuals, we consider eDNA read frequencies to estimate population allele frequencies and therefore conduct genetic analyses that can be implemented at the population level. Although individual-level information is available for Round Goby tissues, we also conducted tissue-based analyses at the population level to allow for direct comparison between eDNA- and tissue-based population genetic analysis. To

compare the efficiency of sampling individual tissues and collecting eDNA samples at a site, we constructed accumulation curves of allelic richness per eDNA sample and sampled individual using `vegan` 2.5–6 (Oksanen et al., 2013). Normalized reads of alleles were combined across triplicate eDNA samples per site and the read frequencies per site were compared to allele frequencies from genotyped tissues using a Pearson's correlation test. From these allele frequencies, we calculated population-level allelic richness, expected heterozygosity, and the number of private alleles. To visualize spatial patterns of genetic variation, we conducted a principal components analysis (PCA) of the centered population allele frequencies in eDNA and tissue samples.

Because eDNA samples do not contain individual genotypes, we calculated pairwise genetic distances using the absolute allele frequency difference among populations (AFD; (Berner, 2019). This metric is an intuitive alternative to F_{ST} that requires only population-level allele frequencies in its estimation and is therefore applicable to our eDNA dataset. Pairwise population estimates of AFD and F_{ST} are highly correlated in Round Goby tissues (Mantel test with 999 permutations; $r = 0.81$, $p < 0.001$). We assessed the similarity between tissue-based and eDNA-based pairwise measures of AFD using Pearson's correlation test and assessed the correlation between AFD and Euclidean geographic distances among eDNA samples using a Mantel test. Last, we used pairwise AFD estimates from allele frequencies estimated from both eDNA and Round Goby tissues to construct unrooted neighbor-joining trees using the method of Saitou and Nei (1987) in `ape` (Paradis, Claude, & Strimmer, 2004).

Results

mtDNA and nuDNA concentration in eDNA

Comparisons of mtDNA and nuDNA concentrations revealed that both qPCR assays amplified Round Goby DNA from environmental samples. The mean amplification efficiency and R^2 of the standard curves of across all runs were 94.6 % ± 6.1 (mean ± 1 *SD*) for mtDNA and 0.995 ± 0.002 for nuDNA. No PCR amplification was measured in any of the negative controls from the field, extraction, or laboratory, indicating no sample contamination was detected. Both assays returned zero copies of Round Goby DNA across all technical and field replicates of eDNA samples taken from Seneca Lake, a site in the Finger Lakes region where Round Gobies have not yet been observed. We exclude this site from further analysis.

We detected Round Goby mtDNA (COI assay) and nuDNA (microsatellite assay) in at least two of the three sample replicates at each of the remaining sites, with mtDNA detected in all eDNA samples (100%) and nuDNA detected in 92% of the eDNA samples. Overall, the concentration of mtDNA was significantly higher than the concentration of nuDNA (type III ANOVA with Satterthwaite's method; $F = 896.54$, $p < 0.001$), with an average of 1,300 ± 1,660 copies/µL of mtDNA and 12 ± 19 copies/µL of nuDNA (Figure 4.1A). The amount of mtDNA in each sample was positively associated with the amount of nuDNA in each sample ($R^2 = 0.47$, $p < 0.001$; Figure 4.1B). In samples where both DNA markers were detected, the average ratio of mtDNA:nuDNA concentration was 196:1, but there was substantial variation in this ratio within and among sites (range 20:1–1,132:1; Figure S4.1).

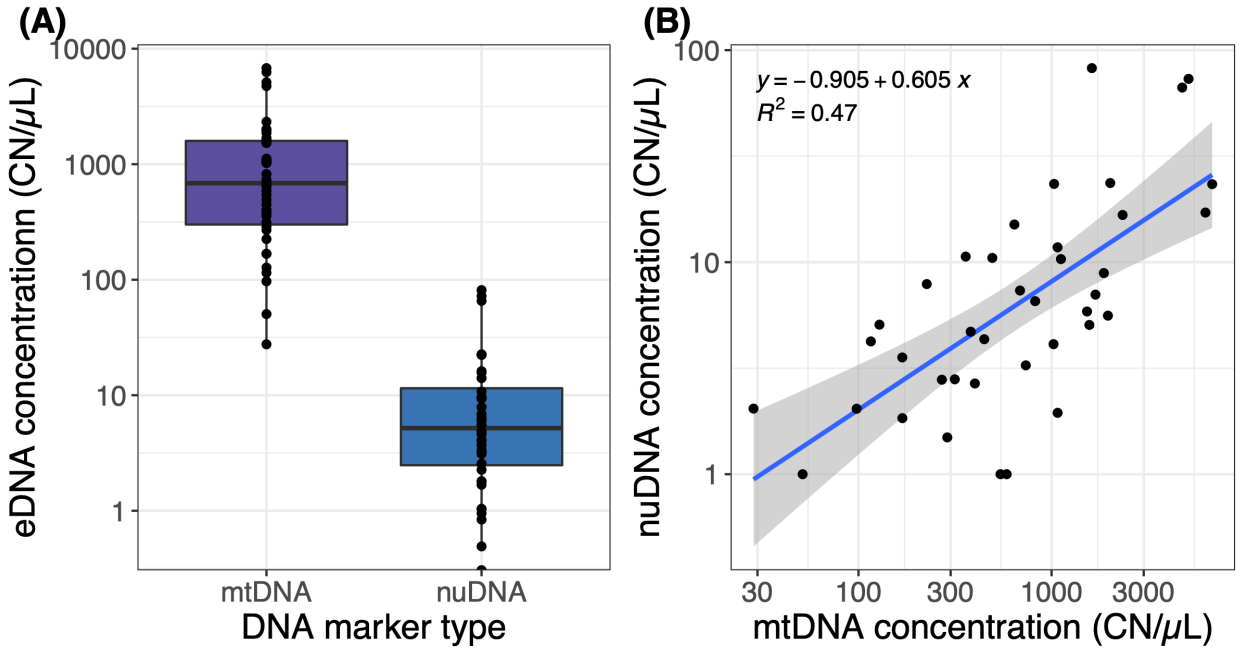

Figure 4.1. Results of the qPCR assays measuring mtDNA and nuDNA concentrations. (A) Boxplot showing the median DNA concentration of each marker type (horizontal line), first and third quartiles (hinges of the box), and 1.5 * the interquartile range (upper/lower whiskers). (B) Correlation between log-transformed mtDNA and nuDNA concentrations. Each dot represents the DNA concentration of a single sample, and the blue line/gray area represents the linear model with a 95% confidence interval.

Tissue-based genetic diversity and population structure

We identified 389 total alleles across 27 loci in the 299 analyzed Round Goby individuals, with an average of 178.9 ± 21.8 alleles per population (mean 6.63 ± 0.81 alleles per locus; Table 4.1). Allelic richness significantly differed among populations (ANOVA; allelic richness: $F = 2.33$, $p = 0.007$), but was not associated with time since invasion ($t = 0.433$, $p = 0.674$). However, the number of private alleles per population was significantly associated with time since invasion, with the highest number of private alleles found at the

oldest populations near the site of original introduction to North America ($t = 2.234$, $p = 0.047$). Pairwise F_{ST} values ranged from 0 to 0.12 and were significantly different from zero in most pairwise comparisons (Figure S4.2). The Mantel test identified a significant relationship between pairwise genetic distance and geographic distance ($r = 0.744$, $p = 0.001$) indicating a pattern of isolation by distance among Round Goby populations.

Table 4.1. Genetic diversity indices of sampled Round Goby populations in North America calculated based on tissue and eDNA samples. Total number of alleles (N_A), average number of alleles per locus (± s.d.; N_{AL}), expected heterozygosity (H_E), and number of private alleles (P_A) are calculated based on genotyped Round Goby tissues and eDNA samples, while observed heterozygosity (H_O) and inbreeding coefficient (F_{IS}) are calculated only based on tissues. Metrics are not shown for sites that were excluded from the eDNA dataset based on low read counts.

			Tissues							eDNA			
Site	**Latitude**	**Longitude**	**N**	**N_A**	**N_{AL}**	**H_O**	**H_E**	**F_{IS}**	**P_A**	**N_A**	**N_{AL}**	**H_E**	**P_A**
Lake Michigan (west)	42.579	-87.815	28	175	6.5 ± 2.7	0.384	0.606	0.407	7	–	–	–	–
Lake Michigan (east)	43.224	-86.339	28	167	6.2 ± 2.8	0.356	0.617	0.45	5	125	4.6 ± 2.8	0.447	24
Lake Huron	43.43	-82.539	28	225	8.3 ± 3.1	0.417	0.700	0.445	20	133	4.9 ± 2.6	0.535	32
Lake St. Clair	42.564	-82.784	14	189	7.0 ± 2.8	0.421	0.698	0.437	15	226	8.4 ± 5.5	0.584	80
Lake Erie (west)	42.08	-83.195	9	152	5.6 ± 2.4	0.483	0.652	0.334	5	–	–	–	–
Lake Erie (east)	42.342	-79.595	13	166	6.1 ± 2.6	0.471	0.696	0.366	1	–	–	–	–
Lake Ontario (west)	43.27	-77.626	28	188	7.0 ± 2.9	0.422	0.677	0.409	6	–	–	–	–
Lake Ontario (east)	43.463	-76.519	28	204	7.6 ± 3.1	0.420	0.676	0.422	4	171	6.3 ± 4.4	0.529	34
Erie Canal	43.119	-77.64	20	176	6.5 ± 2.5	0.431	0.687	0.388	4	–	–	–	–
Cayuga Lake	42.896	-76.75	15	143	5.3 ± 2.0	0.420	0.636	0.397	3	139	5.2 ± 3.2	0.527	27
Cross Lake	43.15	-76.483	18	168	6.2 ± 2.5	0.492	0.676	0.329	3	–	–	–	–
Onondaga Lake	43.115	-76.241	28	196	7.3 ± 3.2	0.447	0.696	0.404	1	155	5.7 ± 3.7	0.516	32
Oneida Lake	43.174	-75.93	28	177	6.6 ± 2.8	0.410	0.654	0.407	4	176	6.5 ± 4.0	0.595	48

Results from the Bayesian assignment tests on Round Goby tissues identified the most likely number of genetic clusters at K = 2 (Figure S4.3). Populations at the periphery (i.e., westernmost and easternmost populations) exhibited the highest average assignment values, with >85% of individuals at sites in Lake Michigan and Lake Huron exceeding membership probabilities of 0.8 to group 1, and >75% of individuals in the Finger Lakes region (Erie Canal, Cayuga Lake, Cross Lake, Onondaga Lake, and Oneida Lake) exceeding membership probabilities of 0.8 to group 2. Sites near the center of the sampled region exhibited a mixture of assignments to the two genetic clusters. Visualization of genetic structure with DAPC further supported the separation of populations along this longitudinal gradient, with distinct groupings of peripheral populations and overlapping central populations (Figure S4.3). Hierarchical AMOVA based on genetic clusters and populations indicated that a significant proportion of the genetic variation could be attributed to differences among genetic clusters (3.97%) and populations (5.37%), although was much lower compared to the amount of variation attributed to within-population (90.66%; Table 4.2).

Table 4.2. Hierarchical analysis of molecular variance (AMOVA) conducted on genotyped Round Goby tissues.

AMOVA groups	d.f.	Sum of squares	Mean squares	Estimated variance	Percentage of total variance	Φ_{ST}	Probability (p)
Among clusters	1	88.25	88.25	0.47	3.97	0.04	0.002
Among populations	11	266.44	24.22	0.63	5.37	0.06	< 0.001
Within populations	272	2896.45	10.65	10.65	90.66	0.09	< 0.001

Detection and quantification of genetic diversity from eDNA

Although qPCR did not detect contamination in any of the negative controls, eDNA sequencing exhibited an average of 17.2 total sequence reads across 12.6 alleles (mean read depth of 1.3 reads per allele) in the negative controls. We therefore removed all alleles with fewer than 2 reads in a sample to account for the observed low levels of contamination.

There was substantial variation in the number of sequencing reads, observed alleles, and successfully amplified loci in eDNA samples within and among sites (Figure S4.4). Due to unreliable microsatellite amplification in some of the eDNA samples, we only retained eDNA samples that exhibited an average allelic richness of ≥ 3 alleles in at least 80% of the loci (≥ 21 loci). This filtering led to the removal of all eDNA samples at six of the 13 sampled sites.

Across the remaining eDNA samples, we detected a total of 543 alleles, including 66.8% of all alleles known from genotyped tissues and 83.7% of all alleles that are present in genotyped tissues at greater than 10% frequency (Figure 4.2). Rarefaction curves of the accumulated number of alleles per sample suggest that the alleles detected with tissue sampling began to saturate in the sampled individuals across each locus and all loci combined (Figure S4.5). However, accumulation curves did not saturate at most loci for eDNA samples, and more sampling likely would have recovered more microsatellite alleles. Average allelic richness and expected heterozygosity was lower in eDNA samples than in genotyped tissues (Table 4.1; Figure S4.6), although this difference was only significant for expected heterozygosity (paired t-test, $t = 8.277$, $p < 0.001$). On the other hand, the number of private alleles was significantly higher based on eDNA samples than tissues (paired t-test, $t = -4.816$, $p = 0.003$; Figure S4.6). None of these summary statistics were significantly correlated between eDNA samples and tissue samples from the same population (i.e., taken at the same

site). The number of private alleles in eDNA samples did not exhibit significant association with time since invasion, as was the case based on genotyped tissues.

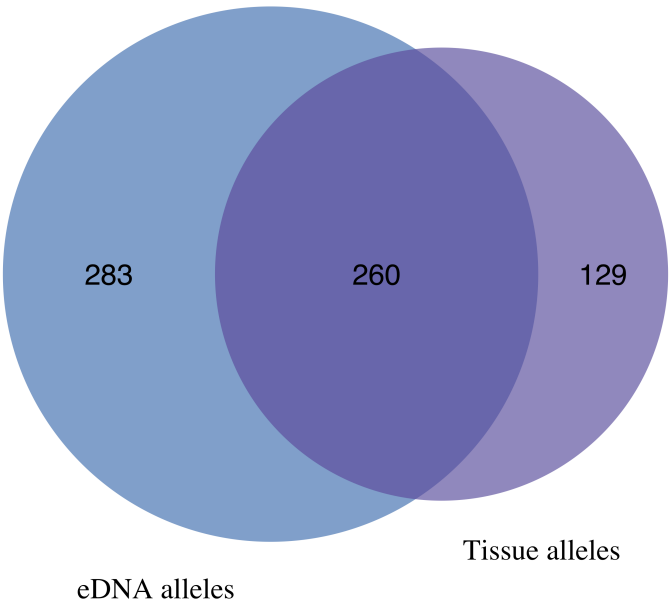

Figure 4.2. Venn diagram of Round Goby microsatellite alleles detected in eDNA samples and genotyped tissues.

Across all sites, allele frequencies from genotyped tissues were positively correlated with eDNA read frequencies ($r = 0.69$, $p < 0.001$), and correlation coefficients per site ranged from 0.59–0.85 (Figure S4.7). PCA of population allele frequencies from genotyped tissues and eDNA samples show congruent patterns of genetic differentiation, where the first principal component axis differentiates populations along the longitudinal gradient (Figure 4.3A–B). Pairwise allele frequency distances (AFD) calculated from eDNA read frequencies was positively correlated with AFD measured from tissue-based allele frequencies ($r = 0.76$, $t = 5.147$, $p < 0.001$), although AFD values were higher for eDNA than for tissues (Figure 4.4).

Neighbor-joining trees based on pairwise population AFD values showed similar patterns of population associations, with the greatest differentiation between the eastern and western sites (Figure 4.3C). This pattern of isolation by distance was further supported in the eDNA data with a Mantel test, which found a significant association between pairwise AFD and geographic distance among all populations ($r = 0.38$, $p = 0.037$).

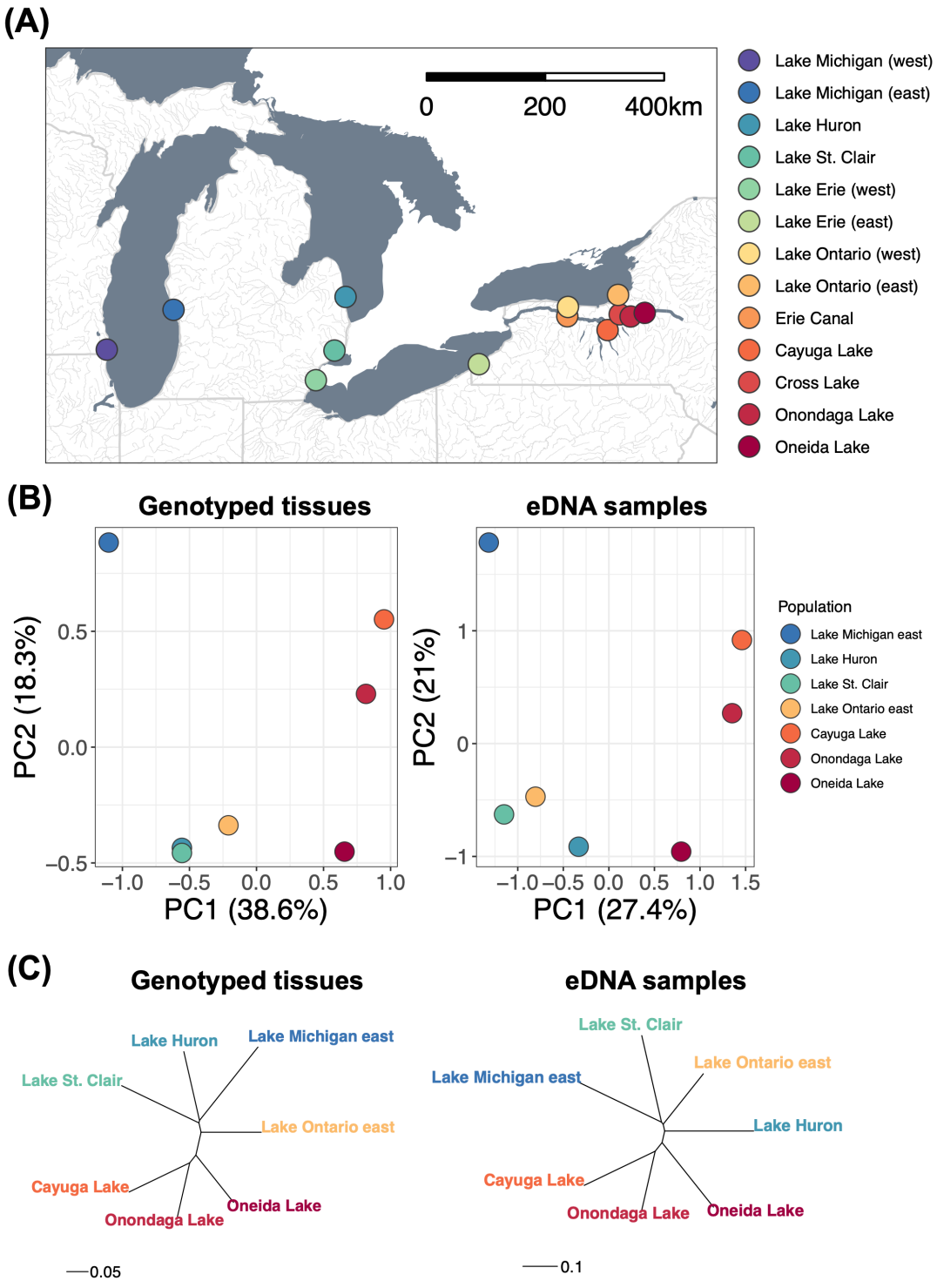

Figure 4.3. (A) Locations sampled for Round Goby tissues and eDNA; (B) PCA of tissue-based allele frequencies (left) and eDNA-based allele frequencies (right); (C) Unrooted

neighbor-joining trees constructed from a matrix of genetic distance using allele frequencies derived from Round Goby tissues (left) and eDNA samples (right).

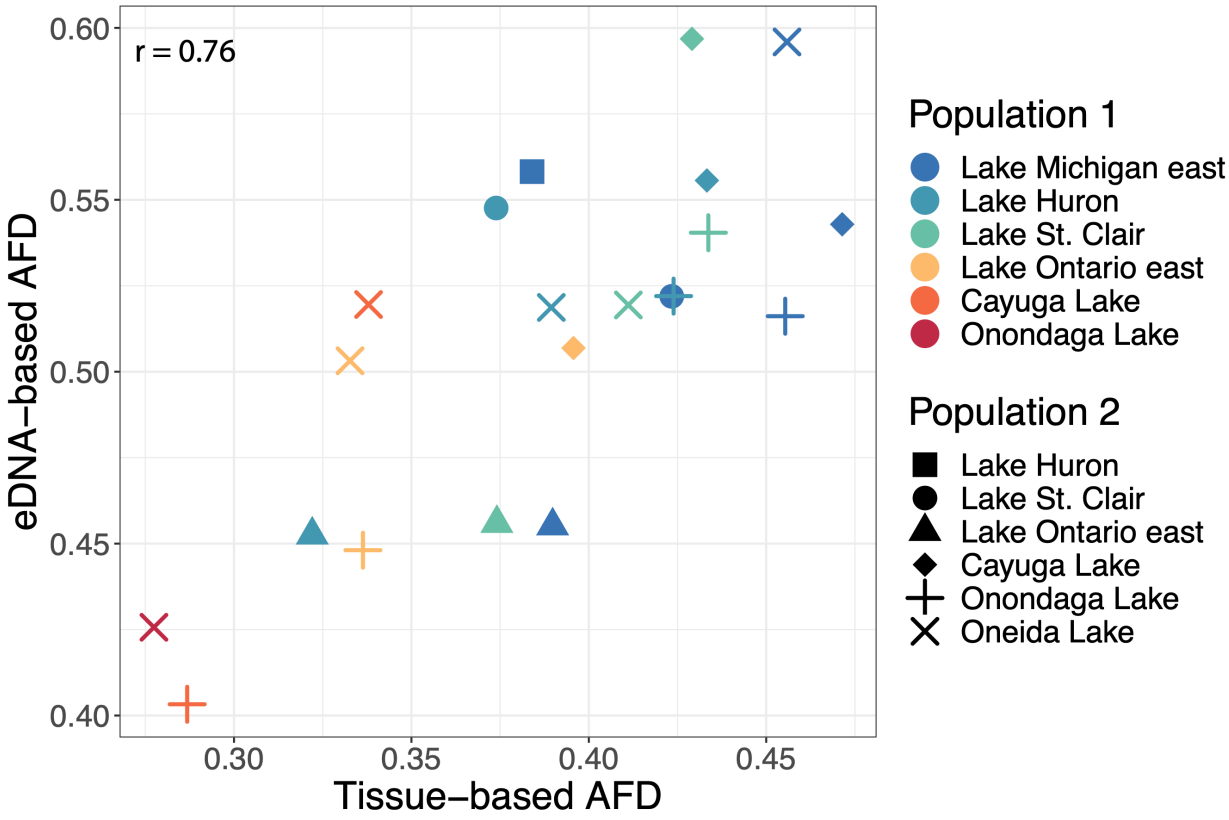

Figure 4.4. Correlation of pairwise AFD values calculated from eDNA read frequencies and tissue allele frequencies. Each point represents a population pair, with color and shape indicating the populations.

Discussion

Environmental DNA is a sensitive and powerful tool for biodiversity assessments, but environmental samples contain much more genetic information than is typically analyzed. In

this study, we used metabarcoding of a microsatellite panel to show that eDNA approaches can be used to detect nuclear genetic variation within species in field settings, making it possible to estimate population allele frequencies and analyze genetic diversity and structure from environmental samples. The amplification of a panel of nuclear genetic markers in natural settings represents a substantial advance toward population-level analyses from eDNA samples, which to date have been limited to analysis of mitochondrial markers (Sigsgaard et al., 2020). Although we found as expected that nuDNA was much less abundant in eDNA samples than mtDNA, we were nonetheless able to amplify at least 80% of the loci in over half of the eDNA samples. The resulting estimates of population allele frequencies and genetic differentiation were highly correlated between eDNA and tissue-based approaches.

Round Gobies were first introduced to North America over 30 years ago, and previous population genetic work identified high levels of diversity at the sites of initial invasion, limited gene flow among North American sites, and slightly higher diversity at the invasion core than at the periphery (Brown & Stepien, 2009; LaRue et al., 2011). We found similarly high levels of genetic diversity and structuring through the analysis of Round Goby tissues, with most populations genetically distinguishable from one another and the primary pattern of divergence following a longitudinal gradient. As in Brown and Stepien (2009), our tissue-based analysis supports the leading-edge hypothesis, with the greatest levels of genetic diversity found at the earliest sites of Round Goby invasion and lower levels of diversity at the expansion edge. We found similar patterns of genetic structuring and differentiation based on analysis of eDNA samples collected at the same sites as Round Goby tissues, with populations structured along the same longitudinal gradient and exhibiting differentiation

from one another. Although the leading-edge model of expansion was not supported in eDNA samples, we did observe a pattern of isolation by distance in both tissue and eDNA samples.

Although we retrieved more total alleles from eDNA (n = 543) than from tissues (n = 389), the number of alleles detected per site was lower in eDNA than in tissues (Table 4.1). Reduced detection of alleles from eDNA might be attributed to low concentrations of nuclear eDNA coupled with small numbers of samples per site leading to reduced detection and incomplete sampling of the genetic diversity of the population. Accordingly, we did not observe plateauing of the accumulation curves of microsatellite alleles across eDNA samples (Figure S4.5), indicating that each eDNA sample provided new genetic information, a trend that was not observed in the tissue samples. Similar accumulation patterns were observed in Sigsgaard et al. (2016), who recovered more haplotypes in eDNA and did not saturate the genetic diversity in the population with their eDNA sampling scheme. Our results show that alleles that are rare in genotyped Round Goby tissues were less likely to be recovered in eDNA samples, likely due to stochastic sampling of low-concentration markers. Similar challenges of recovering rare genetic variants were found in the analysis of mitochondrial haplotypes in Adams et al. (2022), who recovered all common haplotypes but only one rare haplotype of the Blackfoot Pāua (*Haliotis iris)* from seawater samples. Although previous studies have documented approaches to maximizing species detections from eDNA (Bessey et al., 2020; Wilcox et al., 2013), the development and validation of approaches to obtain target DNA in high enough concentrations to reliably amplify rare genetic variants in nuclear markers may be required in future work amplifying nuDNA markers from eDNA. Such optimization could include targeted sampling where the species is known to occur, target capture of specific genes of interest (Jensen et al., 2021), laboratory extraction and

amplification protocols returning higher DNA yields, and/or a greater number or larger-volume samples.

Although we sampled in nearshore habitats where Round Gobies are expected to be in highest abundance during the sampled time period (Andres et al., 2020), we observed very small amounts of nuclear DNA in eDNA samples. Concentrations of nuDNA were much lower than mtDNA, with an average mtDNA:nuDNA ratio of 196:1 across all sampled sites. This is in contrast to previous research investigating the ratios of mtDNA:nuDNA, which targeted multi-copy ribosomal nuDNA markers and found comparable concentrations of mtDNA and nuDNA in eDNA samples (Furtwängler et al., 2018; Jo, Arimoto, Murakami, Masuda, & Minamoto, 2020; Jo, Takao, & Minamoto, 2022; Minamoto et al., 2017). However, much lower concentrations of single-copy nuDNA markers such as microsatellites are expected, as cells contain many more mitochondria than nuclei (Cole, 2016). Low nuDNA template copy numbers may lead to low or unreliable amplification of nuclear genetic markers markers during metabarcoding. Indeed, we observed the lowest qPCR copy numbers in the samples that did not exceed our threshold of the number of loci successfully sequenced (Table S4.1).

In addition to incomplete characterization of population genetic diversity, reduced detection of alleles in eDNA can lead to large differences in allele detections and composition among eDNA samples and can have a significant impact on the resulting genetic analyses. Indeed, we did not observe similar estimates of allelic richness, expected heterozygosity, or number of private alleles per population among eDNA and tissue samples. Due to sporadic allele detections, the number of private alleles detected per site was much higher in eDNA samples, while high frequencies of one or few alleles per eDNA sample led to greatly reduced

estimates of expected heterozygosity. Non-detection of alleles could also explain the absence of support for the leading-edge model in eDNA samples, as this model is based on allelic richness. Genetic distances among populations measured with AFD were also higher for eDNA samples than tissue samples, reflecting inflated genetic distances due to imperfect detection of alleles. Lessons learned from studies of ancient DNA (aDNA; Higuchi, Bowman, Freiberger, Ryder, & Wilson, 1984) and non-invasive genotyping of hair, feathers, or feces (Taberlet & Bouvet, 1992) may be useful for guiding the technical and conceptual challenges faces by population genetic analysis of eDNA samples. For instance, Taberlet, Waits, and Luikart (1999) discuss the potential impacts of errors in non-invasive genotyping research, highlighting the limitations as well as potential solutions to false negative and false positive detections of alleles.

Unlike tissue-based genotyping and population genetic methods, eDNA is a mixture of DNA from potentially many individuals, and eDNA datasets cannot be analyzed using the same theoretical frameworks and analyses developed for genotyped tissues. For instance, many diversity and differentiation parameters rely on knowledge of individual genotypes to assess pairwise measures of genetic identity among individuals. However, any allele detected in eDNA may have been derived from a single or multiple individuals, with no ability to determine the source number of individuals. Furthermore, most population genetic software and statistical frameworks are built for the analysis of genotyped individuals, with few options to analyze genetic parameters with population-level genetic data. Although statistical frameworks and software based on analysis of Pool-seq data (i.e., pooled sequencing of many individuals) have been developed (Hivert, Leblois, Petit, Gautier, & Vitalis, 2018; Kofler, Pandey, & Schlotterer, 2011), these frameworks were developed for bi-allelic SNPs and are

not made to handle multi-allelic microsatellites. Most analysis of Pool-seq data also requires knowledge of the number of individuals sequenced and assume equal contribution of DNA to the pool from all individuals, neither of which is known from eDNA samples (although see Sethi, Larson, Turnquist, Isermann, and Kembel (2019) and Andres et al. (2021) for approaches on estimating the number of contributors to mixed DNA samples). New statistical frameworks that can handle population-level genetic data and account for false positive and false negative alleles, unequal DNA contributions among individuals, and unknown numbers of contributors are therefore needed for robust analysis of eDNA-based population genetic data.

Another challenge in the analysis of intraspecific genetic variation from eDNA is ensuring species specificity of the primers used, as co-amplification of DNA from closely related species can have a strong influence on the conclusions of eDNA research. Although Round Gobies do not have any congeners in North American, the Tubenose Goby (*Proterorhinus semilunaris*) of the Gobiidae family is present in Lake St. Clair and western Lake Erie (Fuller et al., 2022; Kocovsky et al., 2011). Our analysis of eDNA samples from Lake St. Clair revealed the highest numbers of allelic richness and private alleles of all analyzed populations, which could occur due to 1) true high genetic diversity at the site of original invasion; or 2) the co-amplification of Tubenose Goby DNA in these samples. We believe that the first explanation is more likely because we targeted a large panel of loci and rigorously filtered out any alleles from eDNA that did not align with high identity to alleles identified from Round Goby tissues, minimizing the possibility of retaining Tubenose Goby alleles in our analysis. Researchers interested in using eDNA analysis to investigate population genetics should take special care to exclude the possibility of co-amplifying

locally occurring close relatives of the species of interest. Reference databases, particularly for nuclear markers, are unlikely to contain genetic information for all possible close relatives of the target species. For instance, although the Round Goby genome has been sequenced (Adrian-Kalchhauser et al., 2020), the Tubenose Goby genome does not have a reference genome available for comparison. A conservative approach to addressing this issue would be to restrict all genetic variants detected in eDNA to alleles known from tissues (as in Parsons et al., 2018) but this could lead to higher levels of false negatives in eDNA samples.

The Laurentian Great Lakes are among the most heavily invaded ecosystems in the world, with a documented 188 established nonindigenous species, of which over 30% have measurable socioeconomic or environmental impacts (Sturtevant et al., 2019). The most effective strategy for minimizing the impact of invasive species is preventing the introduction and establishment of nonindigenous species altogether, an effort that requires the identification of introduction pathways and efficient monitoring methods (Lodge et al., 2006). Monitoring the presence of nonindigenous species using eDNA approaches can result in higher detection probabilities, allowing for species detections at very low population densities when eradication efforts are less costly and most likely to succeed (Dougherty et al., 2016; Vander Zanden, Hansen, Higgins, & Kornis, 2010). However, if intraspecific genetic variation can be detected in environmental samples, eDNA approaches may also be useful for identifying the geographic source of introduced individuals, elucidating introduction pathways and patterns of subsequent spread. The genetic variation recovered in eDNA samples can also indicate the presence of multiple individuals in the sampled environment, with DNA mixture models able to estimate the number of distinct individuals in a sample (Andres et al., 2021). eDNA therefore provides new opportunities to evaluate population

densities and introduction pathways for nonindigenous species, information that may be critical for developing effective management plans (Lodge et al., 2016; Vander Zanden & Olden, 2008).

In addition to invasive species, eDNA may be useful in the study of intraspecific diversity in species for which tissue collection may be difficult or risks harming the target organism. For instance, marine mammals are a common target in eDNA research (Baker, Steel, Nieukirk, & Klinck, 2018; Dugal et al., 2021; Parsons et al., 2018; Sigsgaard et al., 2016; Székely et al., 2021), at least partially due to the logistical challenges with sampling genetic material of large, potentially deep-diving mammals in remote locations at sea. Obtaining population genetic information from water samples therefore provides an efficient and noninvasive alternative to tissue sampling. Rare or endangered species may also be targeted in population genetic research using eDNA samples, as conventional methods of physically capturing and sampling such species may be ineffective or risk harm to vulnerable species. While this study focused on detecting nuclear genetic diversity in a single species, eDNA samples contain genetic material from many different species. The expansion of this approach to population genetic assessments of multiple species simultaneously is possible (Stat et al., 2017; Weitemier et al., 2021), although this has not yet been demonstrated with nuclear genetic markers.

Biodiversity is declining at alarming rates, with populations of freshwater species declining at a disproportionately faster rate (Reid et al., 2019), and exhibiting particularly severe impacts from invasive species compared to terrestrial systems (Moorhouse & Macdonald, 2015). Conserving freshwater biodiversity requires reliable monitoring of species and populations over space and time, but such data is challenging to collect using

conventional sampling. Environmental DNA provides an opportunity to conduct bioassessments and biomonitoring surveys with improved species detections over other sampling methods, often with lower effort and cost (Bálint et al., 2018; Elbrecht, Vamos, Meissner, Aroviita, & Leese, 2017). With the expansion of eDNA approaches into the realm of population genetics, detailed patterns of genetic diversity within and among species may be easily attainable with the collection of environmental samples alone. Although further efforts are needed to investigate and validate the potential for eDNA metabarcoding to assess within-species genetic diversity, such information will be vital for measuring, predicting, and preventing biodiversity declines across the globe.

Acknowledgements

Funding for this project was provided by the University of Michigan Cooperative Institute for Great Lakes Research (CIGLR), through the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement NA17OAR4320152 (CIGLR contribution), the National Science Foundation (NSF) Coastal SEES program (grant number 1748389), and the Department of Defense (DoD) Strategic Environmental Research and Development Program (SERDP; grant number RC19-1004). We thank Maria Haileselassie, Diana Tapia, George Rogalskyj, Miranda Andres, and Joe Boschert for assistance with collecting specimens in the field. All animals were collected and euthanized according to the Cornell IACUC Animal Care and Use Procedure (ACUP 306.02) and local collecting permits.

References

Adams, C. I., Hepburn, C., Jeunen, G.-J., Cross, H., Taylor, H. R., Gemmell, N. J., . . . Knapp, M. (2022). Environmental DNA reflects common haplotypic variation. *Environmental DNA*.

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G.-J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: can environmental DNA (eDNA) cut it as a population genetics tool? *Genes, 10*(3), 192.

Adrian-Kalchhauser, I., Blomberg, A., Larsson, T., Musilova, Z., Peart, C. R., Pippel, M., . . . Winkler, S. (2020). The round goby genome provides insights into mechanisms that may facilitate biological invasions. *BMC biology, 18*(1), 1-33.

Andres, K. J., Sethi, S. A., Duskey, E., Lepak, J. M., Rice, A. N., Estabrook, B. J., . . . Scofield, A. E. (2020). Seasonal habitat use indicates that depth may mediate the potential for invasive round goby impacts in inland lakes. *Freshwater Biology, 65*(8), 1337-1347.

Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular Ecology, 30*(3), 685-697.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: *Babraham Bioinformatics*, Babraham Institute, Cambridge, United Kingdom.

Baker, C. S., Steel, D., Nieukirk, S., & Klinck, H. (2018). Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. *Frontiers in Marine Science, 5*, 133.

Bálint, M., Nowak, C., Márton, O., Pauls, S. U., Wittwer, C., Aramayo, J. L., . . . Jansen, M. (2018). Accuracy, limitations and cost efficiency of eDNA-based community survey in tropical frogs. *Molecular Ecology Resources, 18*(6), 1415-1426.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology, 13*(4), 729-744.

Balshine, S., Verma, A., Chant, V., & Theysmeyer, T. (2005). Competitive interactions between Round Gobies and logperch. *Journal of Great Lakes Research, 31*(1), 68-77.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Berner, D. (2019). Allele frequency difference AFD–an intuitive alternative to FST for quantifying genetic population differentiation. *Genes, 10*(4), 308.

Bessey, C., Jarman, S. N., Berry, O., Olsen, Y. S., Bunce, M., Simpson, T., . . . Keesing, J. (2020). Maximizing fish detection with eDNA metabarcoding. *Environmental DNA, 2*(4), 493-504.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bronnenhuber, J. E., Dufour, B. A., Higgs, D. M., & Heath, D. D. (2011). Dispersal strategies, secondary range expansion and invasion genetics of the nonindigenous Round Goby, Neogobius melanostomus, in Great Lakes tributaries. *Molecular Ecology, 20*(9), 1845-1859.

Brown, J. E., & Stepien, C. A. (2009). Invasion genetics of the Eurasian round goby in North America: tracing sources and spread patterns. *Molecular Ecology, 18*(1), 64-79. doi:10.1111/j.1365-294X.2008.04014.x

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583.

Cole, L. W. (2016). The evolution of per-cell organelle number. *Frontiers in Cell and Developmental Biology, 4*, 85.

Crane, D. P., & Einhouse, D. W. (2016). Changes in growth and diet of smallmouth bass following invasion of Lake Erie by the Round Goby. *Journal of Great Lakes Research, 42*(2), 405-412.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., . . . Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish Orconectes rusticus at low abundances. *Journal of Applied Ecology, 53*(3), 722-732.

Dray, S., & Dufour, A.-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software, 22*, 1-20.

Dugal, L., Thomas, L., Reinholdt Jensen, M., Sigsgaard, E. E., Simpson, T., Jarman, S., . . . Meekan, M. (2021). Individual haplotyping of whale sharks from seawater environmental DNA. *Molecular Ecology Resources*.

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution, 8*(10), 1265-1275.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology, 14*(8), 2611-2620.

Evans, N. T., Shirey, P. D., Wieringa, J. G., Mahon, A. R., & Lamberti, G. A. (2017). Comparative cost and effort of fish distribution detection via environmental DNA analysis and electrofishing. *Fisheries, 42*(2), 90-99.

Fediajevaite, J., Priestley, V., Arnold, R., & Savolainen, V. (2021). Meta-analysis shows that environmental DNA outperforms traditional surveys, but warrants better reporting standards. *Ecology and Evolution, 11*(9), 4803-4815.

Freedman, J. A., Fuller, P., Benson, A., Maynard, E., Neilson, M. E., Larson, J., . . . Sturtevant, R. (2022, 4/26/2022). *Neogobius melanostomus* (Pallas, 1814): U.S. Geological Survey, Nonindigenous Aquatic Species Database. Retrieved from https://nas.er.usgs.gov/queries/factsheet.aspx?SpeciesID=713

Fuller, P., L. Nico, E. Maynard, M. Neilson, J. Larson, T.H. Makled, . . . Sturtevant, R. (2022). *Proterorhinus semilunaris* (Heckel, 1837). Retrieved from https://nas.er.usgs.gov/queries/factsheet.aspx?SpeciesID=714

Furtwängler, A., Reiter, E., Neumann, G. U., Siebke, I., Steuri, N., Hafner, A., . . . Krause, J. (2018). Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. *Scientific Reports, 8*(1), 1-8.

George, S. D., Baldigo, B. P., Rees, C. B., Bartron, M. L., & Winterhalter, D. (2021). Eastward expansion of Round Goby in New York: assessment of detection methods and current range. *Transactions of the American Fisheries Society, 150*(2), 258-273.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes, 5*(1), 184-186.

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature, 312*(5991), 282-284.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from Pool-seq data. *Genetics, 210*(1), 315-330.

Janssen, J., & Jude, D. J. (2001). Recruitment Failure of Mottled Sculpin *Cottus bairdi* in Calumet Harbor, Southern Lake Michigan, Induced by the Newly Introduced Round Goby *Neogobius melanostomus*. *Journal of Great Lakes Research, 27*(3), 319-328. doi:10.1016/s0380-1330(01)70647-8

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., . . . Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular Ecology Resources, 21*(3), 690-702.

Jo, T., Arimoto, M., Murakami, H., Masuda, R., & Minamoto, T. (2020). Estimating shedding and decay rates of environmental nuclear DNA with relation to water temperature and biomass. *Environmental DNA, 2*(2), 140-151.

Jo, T., Takao, K., & Minamoto, T. (2022). Linking the state of environmental DNA to its application for biomonitoring and stock assessment: Targeting mitochondrial/nuclear genes, and different DNA fragment lengths and particle sizes. *Environmental DNA, 4*(2), 271-283.

Johnson, T. B., Bunnell, D. B., & Knight, C. T. (2005). A potential new energy pathway in central Lake Erie: the Round Goby connection. *Journal of Great Lakes Research, 31*, 238-251.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics, 11*(1), 1-15.

Jude, D. J., Reider, R. H., & Smith, G. R. (1992). Establishment of Gobiidae in the Great Lakes Basin. *Canadian Journal of Fisheries and Aquatic Sciences, 49*(2), 416-421. doi:10.1139/f92-047

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ, 2*, e281.

Kocovsky, P., Tallman, J., Jude, D., Murphy, D., Brown, J., & Stepien, C. (2011). Expansion of tubenose gobies *Proterorhinus semilunaris* into western Lake Erie and potential effects on native species. *Biological Invasions, 13*(12), 2775-2784.

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics, 27*(24), 3435-3436. doi:10.1093/bioinformatics/btr589

LaRue, E. A., Ruetz, C. R., Stacey, M. B., & Thum, R. A. (2011). Population genetic structure of the round goby in Lake Michigan: implications for dispersal of invasive species. *Hydrobiologia, 663*(1), 71-82.

Lodge, D. M., Simonin, P. W., Burgiel, S. W., Keller, R. P., Bossenbroek, J. M., Jerde, C. L., . . . Zhang, H. (2016). Risk analysis and bioeconomics of invasive species to inform policy and management. *Annual Review of Environment and Resources. 41: 17.1-17.36., 41*, 17.11-17.36.

Lodge, D. M., Williams, S., MacIsaac, H. J., Hayes, K. R., Leung, B., Reichard, S., . . . Andow, D. A. (2006). Biological invasions: recommendations for US policy and management. *Ecological Applications, 16*(6), 2035-2054.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp Cyprinus carpio. *Molecular Ecology Resources, 17*(2), 324-333.

Moorhouse, T. P., & Macdonald, D. W. (2015). Are invasives worse in freshwater than terrestrial ecosystems? *Wiley Interdisciplinary Reviews: Water, 2*(1), 1-8.

Nathan, L. R., Jerde, C. L., Budny, M. L., & Mahon, A. R. (2015). The use of environmental DNA in invasive species surveillance of the Great Lakes commercial bait trade. *Conservation Biology, 29*(2), 430-439.

Nevers, M. B., Byappanahalli, M. N., Morris, C. C., Shively, D., Przybyla-Kelly, K., Spoljaric, A. M., . . . Roseman, E. F. (2018). Environmental DNA (eDNA): A tool for quantifying the abundant but elusive round goby (*Neogobius melanostomus*). *PLoS One, 13*(1), e0191720. doi:10.1371/journal.pone.0191720

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., . . . Oksanen, M. J. (2013). Package 'vegan'. *Community ecology package, version*, *2*(9), 1-295

Padgham, M., & Sumner, M. D. (2021). geodist: Fast, dependency-free geodesic distance calculations. *R package version 0.0.7, 4*.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics, 20*(2), 289-290.

Parsons, K. M., Everett, M., Dahlheim, M., & Park, L. (2018). Water, water everywhere: Environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science, 5*(8), 180537.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T., . . . Ormerod, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews, 94*(3), 849-873.

Ricciardi, A., & MacIsaac, H. J. (2000). Recent mass invasion of the North American Great Lakes by Ponto–Caspian species. *Trends in Ecology & Evolution, 15*(2), 62-65.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425.

Sepulveda, A. J., Schabacker, J., Smith, S., Al-Chokhachy, R., Luikart, G., & Amish, S. J. (2019). Improved detection of rare, endangered and invasive trout in using a new large-volume sampling method for eDNA capture. *Environmental DNA, 1*(3), 227-237.

Sethi, S. A., Larson, W., Turnquist, K., Isermann, D., & Kembel, S. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution, 10*(1), 109-119. doi:10.1111/2041-210x.13079

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Moller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA-Current status and future perspectives. *Evolutionary Applications, 13*(2), 245-262. doi:10.1111/eva.12882

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., . . . Thomsen, P. F. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution, 1*(1), 4. doi:10.1038/s41559-016-0004

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S. T., . . . Yu, D. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution, 8*(5), 635-645. doi:10.1111/2041-210x.12683

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., . . . Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports, 7*(1), 1-11.

Stepien, C. A., & Tumeo, M. A. (2006). Invasion genetics of Ponto-Caspian gobies in the Great Lakes: a 'cryptic'species, absence of founder effects, and comparative risk analysis. *Biological Invasions, 8*(1), 61-78.

Sturtevant, R., Mason, D., Rutherford, E., Elgin, A., Lower, E., & Martinez, F. (2019). Recent history of nonindigenous species in the Laurentian Great Lakes; An update to Mills et al., 1993 (25 years later). *Journal of Great Lakes Research, 45*(6), 1011-1035.

Székely, D., Corfixen, N. L., Mørch, L. L., Knudsen, S. W., McCarthy, M. L., Teilmann, J., . . . Olsen, M. T. (2021). Environmental DNA captures the genetic diversity of bowhead whales (*Balaena mysticetus*) in West Greenland. *Environmental DNA, 3*(1), 248-260.

Taberlet, P., & Bouvet, J. (1992). Bear conservation genetics. *Nature, 358*(6383), 197-197.

Taberlet, P., Waits, L. P., & Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution, 14*(8), 323-327.

Taraborelli, A. C., Fox, M. G., Johnson, T. B., & Schaner, T. (2010). Round goby (*Neogobius melanostomus*) population structure, biomass, prey consumption and mortality from predation in the Bay of Quinte, Lake Ontario. *Journal of Great Lakes Research, 36*(4), 625-632. doi:10.1016/j.jglr.2010.07.011

Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., . . . Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology, 21*(11), 2565-2573.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation, 183*, 4-18. doi:10.1016/j.biocon.2014.11.019

Truett, G. E., Heeger, P., Mynatt, R., Truett, A., Walker, J., & Warman, M. (2000). Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). *Biotechniques, 29*(1), 52-54.

Uchii, K., Doi, H., & Minamoto, T. (2016). A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Molecular Ecology Resources, 16*(2), 415-422.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., . . . Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Vander Zanden, M. J., Hansen, G. J., Higgins, S. N., & Kornis, M. S. (2010). A pound of prevention, plus a pound of cure: early detection and eradication of invasive species in the Laurentian Great Lakes. *Journal of Great Lakes Research, 36*(1), 199-205.

Vander Zanden, M. J., & Olden, J. D. (2008). A management framework for preventing the secondary spread of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences, 65*(7), 1512-1522.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 1358-1370.

Weitemier, K., Penaluna, B. E., Hauck, L. L., Longway, L. J., Garcia, T., & Cronn, R. (2021). Estimating the genetic diversity of Pacific salmon and trout using multigene eDNA metabarcoding. *Molecular Ecology, 30*(20), 4970-4990.

Wilcox, T. M., McKelvey, K. S., Young, M. K., Jane, S. F., Lowe, W. H., Whiteley, A. R., & Schwartz, M. K. (2013). Robust detection of rare species using environmental DNA: the importance of primer specificity. *PLoS One, 8*(3), e59520.

CHAPTER 5

ESTIMATING ROUND GOBY ABUNDANCE USING ADVANCED SURVEY METHODS: A COMPARISON OF AUV IMAGING AND EDNA ASSESSMENTS

Abstract

Accurate estimates of invasive species abundance are critical for evaluating their status, predicting their impacts, and developing management plans to combat their spread. However, assessing the status of bottom-dwelling aquatic species in large lakes remains a challenge. Environmental DNA (eDNA) approaches may be used to estimate species abundance using the quantity of DNA or the amount of genetic variation measured in an environmental sample, but the accuracy of these different abundance estimation frameworks have not yet been compared in natural environments. We test multiple eDNA approaches to determine if DNA marker type, sample volume, and quantification approach (DNA concentration or genetic diversity) influence the relationship between Round Goby (*Neogobius melanostomus*) abundance estimates derived from eDNA and visual observation from AUV images in Lake Huron. We found that DNA concentration assessed with qPCR of a mitochondrial marker from small-volume (0.5 L) eDNA samples is significantly associated with AUV-based abundance estimates, while the same approach using large-volume (100 L) samples shows no association with AUV images. Due to difficulties associated with low DNA concentrations and primer specificity, we were unable to assess genetic diversity-based approaches to estimating Round Goby abundance using eDNA. Although improvements to eDNA-based abundance estimates are possible and required for robust species assessments with eDNA, we

suggest careful consideration of the potential issues arising from different genetic markers, sampling protocols, and methodological approaches in eDNA research.

Introduction

Biological invasions are among the leading causes of freshwater biodiversity declines worldwide and can have significant ecological, evolutionary, and economic impacts on invaded ecosystems (Pimentel, Zuniga, & Morrison, 2005; Sala et al., 2000). Reliable quantitative information on the abundance of invasive species is necessary for understanding their ecological success and impacts and developing plans for their management, but such information is seldom available and often difficult to obtain (Bradley et al., 2018). Abundance information on invasive species may be particularly useful in the early stages of an invasion to assess the risk of establishment, as high abundance (i.e., propagule pressure) is linked to higher establishment probability (Kolar & Lodge, 2001). Furthermore, because high abundance is associated with larger species ranges (Brown, 1984; Gaston & Lawton, 1990), high species abundance at the range edge of an invasive species may suggest that further range expansion is likely to follow (Arim, Abades, Neill, Lima, & Marquet, 2006). The impacts of invasive species may also vary depending on their density, as some studies have found little or no negative ecological impacts of invasives when their densities are similar to that of native species (Aday, 2007; Douglas & O'Connor, 2003) and greater impacts when invasive species densities are high (Elgersma & Ehrenfeld, 2011; Jackson, Ruiz-Navarro, & Britton, 2015). Thus, assessing the degree of a species invasion and predicting the potential threat requires quantification of the abundance of invasive species.

The Round Goby (*Neogobius melanostomus*) is a prolific Great Lakes invader that was originally introduced to the St. Clair River (Michigan, USA) in 1990 via ballast water from the Ponto-Caspian region (Jude, Reider, & Smith, 1992). Since their introduction, Round Gobies have rapidly spread throughout the Great Lakes region and have been linked to

population declines of native fish and invertebrates (Freedman et al., 2022; Janssen & Jude, 2001; Krakowiak & Pennuto, 2008; Lauer, Allen, & McComish, 2004; Steinhart, Marschall, & Stein, 2004). However, they also play an important role in the food web as one of few consumers of invasive dreissenid mussels (Ricciardi & MacIsaac, 2000) and as a plentiful prey for many important fisheries species, transferring energy and nutrients in the benthos to higher trophic levels (Johnson, Bunnell, & Knight, 2005; King, Ray, & Stanford, 2006). Although efficient approaches to assess Round Goby abundance are essential for understanding their ecological impacts and potential management, reliable abundance estimates using conventional sampling gears such as trawls, traps, and nets are difficult to obtain due to their preference for complex benthic habitats (Leino & Mensinger, 2016; Steingraeber, Runstrom, & Thiel, 1996). Surveying gobies with benthic images or direct observation is therefore more effective for assessing their density (Johnson, Allen, Corkum, & Lee, 2005), although even these approaches may be limited due to imperfect visual detections as a result of species camouflage and vegetated habitats (Andres et al., 2020). Therefore, a sensitive approach for assessing Round Goby abundance is needed.

Environmental DNA (eDNA) is an efficient tool to census species using DNA collected and extracted from an environmental sample (Taberlet, Bonin, Zinger, & Coissac, 2018). While eDNA is well established for detecting species presence and biodiversity (Deiner et al., 2017; McElroy et al., 2020; Thomsen & Willerslev, 2015), the utility of eDNA for predicting species abundance remains relatively understudied, particularly in natural systems (Yates, Fraser, & Derry, 2019). To date, most efforts to assess species abundance with eDNA have focused on correlating ambient eDNA concentration with numerical abundance or biomass. Such efforts have been met with moderate success in natural

environments (Doi et al., 2017; Pilliod, Goldberg, Arkle, & Waits, 2013), yet many questions and concerns about the accuracy and interpretation of these results remain, potentially hindering the application of this approach in practice (Yates et al., 2019). This is because the relationship between the amount of eDNA collected in a sample and organismal abundance in the sampled environment may be confounded due to several biotic and abiotic factors influencing DNA production and loss rates (Barnes & Turner, 2015; Strickler, Fremier, & Goldberg, 2015).

For instance, the production rate of eDNA can vary among individuals as a function of their size, behavior, or metabolism, obscuring the relationship between eDNA concentration and organism abundance (Doi et al., 2017; Pilliod et al., 2013). Additionally, although the vast majority of eDNA studies target the mitochondrial genome, mitochondrial DNA (mtDNA) copy number per cell can be highly variable among tissue types within an individual (Robin & Wong, 1988), meaning that different cell types can produce varying amounts of mtDNA. Environmental stochasticity and patchily distributed species can also influence eDNA dynamics, resulting in variable DNA concentrations that cannot be explained by variation in species abundance (Shogren et al., 2017; Troth, Sweet, Nightingale, & Burian, 2021). Understanding and accounting for the sources of DNA concentration variation in environmental samples is therefore a requisite step in obtaining robust estimates of species abundance based on eDNA abundance.

This problem may be overcome, however, by estimating abundance using the amount of genetic diversity detected in environmental samples. In contrast to estimating species abundance from eDNA concentrations, this approach estimates the number of genetic contributors to eDNA samples by comparing the alleles detected in an eDNA sample to

estimated population-level allele frequencies (Andres, Sethi, Lodge, & Andrés, 2021; Sethi, Larson, Turnquist, Isermann, & Kembel, 2019). Andres et al. (2021) demonstrated accurate genetic diversity-based estimates of the number of Round Goby individuals using eDNA samples from experimental mesocosms containing densities of up to 10 individuals. Simulations of DNA mixtures using different marker types support the potential to estimate up to 100 individuals in a sample provided all genetic variants can be detected, particularly those at low frequencies (Chapter 2 of this dissertation; Figure 2.3). However, genetic diversity-based abundance estimates from eDNA have not yet been validated in natural field settings or compared to estimates of DNA concentration.

Additionally, the influence of different target DNA amplification sequences (nuclear vs. mitochondrial) on estimates of species-specific abundance from eDNA samples has not been demonstrated. Because mtDNA is present in a higher copy number per cell than single-copy nuclear DNA (nuDNA), it is expected to be more abundant in environmental samples and particle concentration should therefore be higher, resulting in more accurate species detections. However, nuDNA sequences per cell do not vary among source tissue types, so nuDNA markers may be less sensitive to eDNA production variability within individuals. To date, no study has identified which target sequence provides the most robust predictions of abundance from eDNA in field settings.

Here we develop novel eDNA-based metrics of Round Goby abundance and compare them to image-based abundance estimates in a field sampling effort. We combine eDNA sampling with benthic images to examine Round Goby abundance estimates using a suite of eDNA-based abundance indices: (1) mtDNA particle concentration; (2) nuDNA particle concentration; and (3) mtDNA genetic diversity. To determine the optimal approach for

estimating Round Goby abundance from eDNA samples, the correlation between each of these detection-corrected metrics of Round Goby abundance from eDNA and benthic images will be assessed. We also assess the influence of sample volume on each of these metrics, with small-volume and large-volume samples taken at each site. We hypothesize that the concentration of DNA particles and the amount of genetic diversity in eDNA samples reflect the local abundance of the target species, but that biotic and abiotic factors introduce different sources and degrees of variation in abundance estimates depending on the eDNA approach used. Such variation is expected to influence the association between eDNA- and benthic image-derived abundance indices. Specifically, because eDNA production varies within and among individuals, we predict that eDNA abundance estimates based on genetic diversity will demonstrate greater accuracy and precision over abundance estimates using eDNA concentration. Moreover, because nuDNA is less sensitive to source tissue types, we predict that nuDNA markers will better reflect the abundance of Round Gobies over mtDNA markers in eDNA samples.

Materials and methods

We sampled 15 sites that spanned the western shore of the main basin of Lake Huron, USA (Figure 5.1). With a surface area of 59,590 km^2 and maximum depth of 229 m, Lake Huron is one of the largest lakes on Earth. It supports a diverse species assemblage including several economically and ecologically important fishery species (Brown, Ebener, & Gorenflo, 1999). Round Gobies were first documented in Lake Huron 1994 and have since been recorded throughout most areas of the lake (Riley et al., 2008; Schaeffer, Bowen, Thomas, French III, & Curtis, 2005). Some piscivorous species have shifted their diets to consume

Round Gobies, which now comprise over 96% of the prey consumed by commercially important Lake Whitefish (*Coregonus clupeaformis*; Pothoven & Madenjian, 2013). Annual surveys by the United States Geological Survey (USGS) aim to assess the status and trends of Lake Huron fisheries and prey fish communities through a combination of trawling and acoustics (Hondorp, O'Brien, Esselman, & Roseman, 2022); advanced survey methods to detect lake-wide productivity and species abundance are under development (Bennion, Warner, Esselman, Hobson, & Kieft, 2019). However, robust and unbiased techniques are still needed to properly characterize the status, trends, and impacts of benthic species including Round Gobies (Riley et al., 2008; Stockwell, Yule, Gorman, Isaac, & Moore, 2006).

Abundance estimates with AUV images

We assessed Round Goby abundance in Lake Huron using images obtained from an L3Harris-Ocean Server Iver3 autonomous underwater vehicle (AUV) modified to carry a 9.3 megapixel color camera. The AUV was programmed to fly at a constant altitude of 1.75 m over the lakebed taking a series of overlapping benthic images along a transect on the lake bottom. At each of fifteen locations along the U.S. side of Lake Huron, the AUV was deployed in nearshore environments to gather 10 km photo transects. Since the images were overlapping, for each transect we first unpacked every 25th image to avoid duplicate counting of fishes. Then, we randomly selected 800 images and applied a usability algorithm to drop any unusable overexposed or underexposed images. Then, for fish detection and abundance estimation, we applied our previously trained deep convolutional neural network-based algorithm named RetinaNet (Lin, Goyal, Girshick, He, & Dollár, 2017). This object detection model was previously trained on more than 35,000 AUV images from Lake Michigan

gathered in 2020 to detect fishes in images by placing bounding boxes over regions of the image with potential Round Goby individuals present. Each bounding box is attributed with an associated probability that a Round Goby is present within the box.

To maximize fish detections in the images from each transect, we applied the trained model with a relatively low probability of Round Goby occurrence (= 0.30) and sent the detected bounding boxes with probable fish in them to be labeled by human observers contracted (from Innodata Inc.) to note whether fish were truly present or absent in each tile. Human observers were trained to identify the body shape of Round Goby, which is the only fish captured in any appreciable density in the 0–30 m depth zone with that morphology. Image area (m^2) was estimated by attributing each pixel with a size and multiplying by the image pixel dimensions. Pixel sizes were estimated using the AUV-measured altitude (using an integrated Teledyne RDI Explorer doppler velocity log) at which each image was taken, the camera focal length, and a correction for underwater magnification caused by light refraction between water and the air behind the lens (Adolfson & Berghage, 1974). We then used fish detections per image to calculate image-wise and transect-level numeric density (fish/m^2) per image and transect. We calculated average and standard deviation of fish density across all labeled photos, including those with no fish present.

Table 5.1. Sampling information on each site in the study, including location, sampling date, number of usable AUV images of 800 randomly selected images, number of AUV images with confirmed fish presence, transect density (fish/m^2) and standard deviation, water depth of triplicate small-volume eDNA samples, and volume of water filtered for large-volume samples.

					AUV information				eDNA information	
	Site	**Latitude**	**Longitude**	**Date**	**Usable images**	**Images w/ fish**	**Density (fish/m^2)**	**SD**	**Sample depth (m)**	**Volume (L)**
1	St. Martin Bay	46.004	-84.556	9/16/21	800	0	0	0	5.4, 4.8, 4.7	100
2	St. Ignace	45.883	-84.720	8/11/21	797	298	0.464	0.74	4.7, 5.3, 4.8	–
3	Cheboygan	45.732	-84.391	9/17/21	764	63	0.080	0.28	5.2, 4.7, 5.0	100.03
4	Hammond Bay	45.593	-84.151	8/28/21	0	–	–	–	5.5, 4.3, 5.5	98.62
5	Rogers City	45.430	-83.819	9/19/21	648	241	0.533	0.88	–	99.99
6	Adams Point	45.408	-83.707	9/20/21	794	269	0.409	0.7	–	99.97
7	Presque Isle	45.342	-83.499	9/21/21	800	140	0.231	4.05	–	99.97
8	Alpena	45.036	-83.218	8/29/21	800	237	0.304	0.53	5.1, 5.0, 5.2	100.02
9	Harrisville	44.757	-83.293	9/24/21	0	–	–	–	5.1, 4.8, 4.8	28.68
10	Oscoda	44.499	-83.303	9/25/21	0	–	–	–	5.1, 5.4, 5.3	26
11	Tawas	44.154	-83.559	9/27/21	0	–	–	–	4.9, 5.1, 5.1	51.43
12	Port Austin	44.013	-83.093	9/26/21	0	–	–	–	5.2, 4.9, 5.5	28.34
13	Harbor Beach	43.906	-82.667	8/26/21	788	358	0.790	1.59	4.5, 5.9, 5.4	100.99
14	Lexington	43.307	-82.525	8/25/21	798	415	0.680	0.94	5.1, 4.7, 4.4	100.56
15	Port Huron	43.174	-82.498	8/24/21	800	431	1.008	8.27	5.0, 5.1, 5.2	101.13

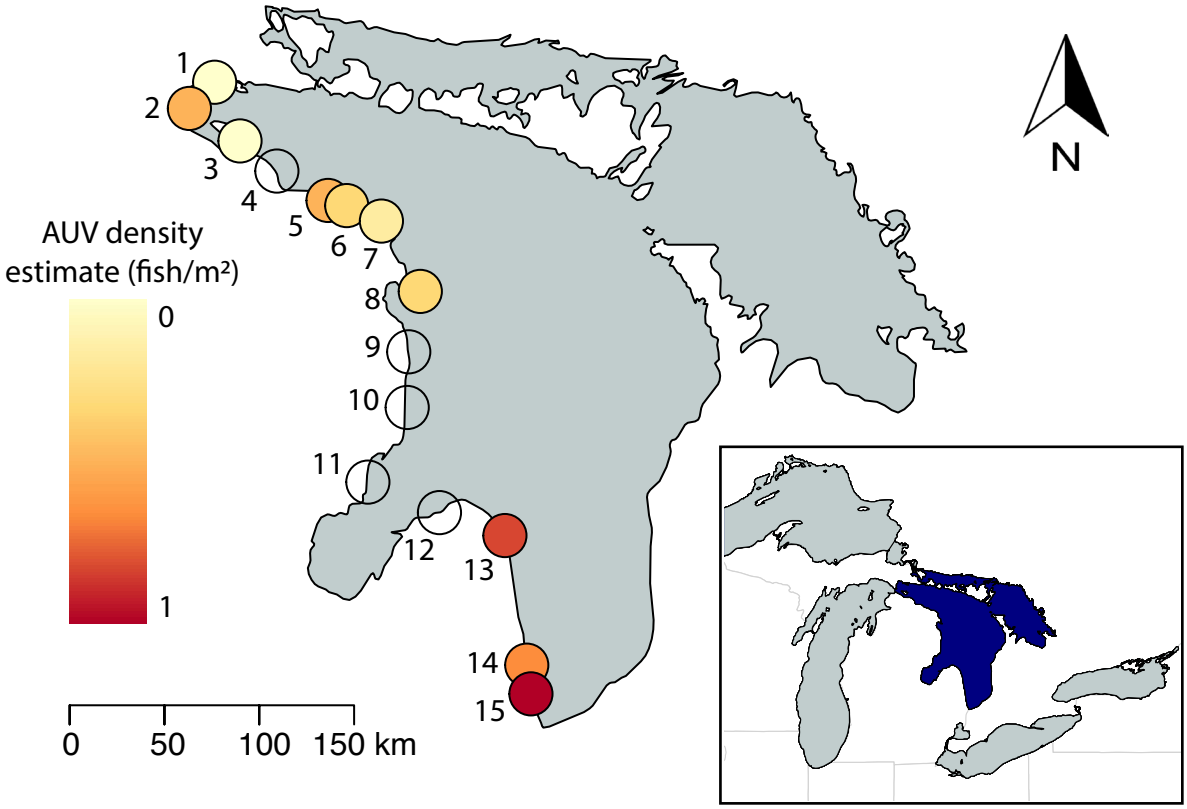

Figure 5.1. Geographical locations of the sampled sites in Lake Huron. Each sampled site is labeled with a corresponding number from Table 5.1. Colors of each point represents the density (fish/m^2) assessed with AUV benthic images, with hollow points representing sites where no usable AUV images were obtained.

eDNA collection and extraction

All eDNA sampling and processing protocols adhered to stringent field and laboratory practices to minimize and detect contamination including sterilizing all reusable equipment,

physical separation of pre- and post-PCR laboratory spaces, and collecting negative controls at each stage of the experiment that were amplified and sequenced to monitor possible contaminants (Goldberg et al., 2016). We collected small-volume and large-volume eDNA samples from each site concurrently with AUV deployment (Table 5.1). During this time of year, the thermocline intersects the lakebed at approximately 20–30 m depth (Rao & Schwab, 2007), and all sampling was completed in depths shallower than the thermocline to maximize homogeneity of the water column. Although Round Gobies are a benthic species, previous research found no difference in Round Goby eDNA concentrations in samples taken from surface waters and just above the lake bottom in the summer months (Nevers et al., 2018). We therefore collected all eDNA samples from surface waters.

Small-volume sampling included six samples collected in pairs approximately 50 m apart, with all sampling occurring from the bow of a forward moving boat, where the water column was 5 m water deep. Each water sample was collected in a sterilized 250 mL wide-mouth Nalgene container (Thermo Fisher Scientific, USA) 0.3 m beneath the water's surface. A disposable 50 mL Luer Lock sterile syringe (Hapool medical Technology) was used to draw water from each collection bottle and pass it through a 1 μm pore size cellulose nitrate filter (25 mm diameter, Whatman, GE Healthcare, USA) contained in a 25 mm Swinnex Filter Holder (Sigma-Aldrich, USA) until 250 mL of water was filtered. Filter membranes were folded with sterile forceps and stored in 700 μL Longmire's buffer (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5% SDS) for up to two weeks at room temperature until long-term storage at -20ºC. To reduce sample processing time, the two filters from each pair of samples were stored together in each microcentrifuge tube and processed as a single sample, resulting in three independent small-volume samples per site, with each sample containing eDNA from

0.5 L of filtered water. Negative controls were also taken at each sampling site by filling a 250 mL Nalgene container with distilled water and filtering it on site. DNA from all small-volume eDNA samples and negative controls were extracted using DNeasy Blood and Tissue extraction kits (Qiagen Inc.) with an optimized protocol for eDNA as described in Chapters 3 and 4 of this dissertation (Spens et al., 2017) and eluted in 100 uL buffer. Due to COVID-induced delays in equipment delivery, we were unable to collect small-volume samples at three of the 15 sites (Table 5.1).

Large-volume eDNA samples were collected at 14 sites using a Smith-Root eDNA Sampler, a fully integrated sampling system equipped with a backpack pump and sensor to monitor pump pressure, flow rate, and filter volume (Thomas, Howard, Nguyen, Seimon, & Goldberg, 2018). We modified the sampling rod with sterile flexible tubing, reducers, and PVC pipe to accommodate Envirocheck HV filters (1 μm, Pall Corporation, USA) capable of filtering over 100 L of water (Figure S5.1). Water was pumped from 0.3 m below the surface at a flow rate of 0.8 L/min until 100 L of water was filtered. Because only a single large-volume eDNA sample was collected per site, we collected the large volume samples in increments: we filtered 25 L of water at four locations that differed in water column depth (5, 10, 15, and 20 m depth), all of which were above the thermocline. Four of the 14 large-volume samples were collected soon after a major storm event, and suspended sediments in the water column prevented filtration of the full 100 L of water (Table 5.1). Upon completion of water filtration, we ran the pump until all water was emptied from the filtration capsule and immediately added 50 mL preservation buffer (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5% N-lauroyl sarcosine). Filter capsules were stored in the dark at room temperature for up to two weeks until storage at -20ºC. DNA extraction of large-volume eDNA samples followed

a modified protocol of DNeasy Blood and Tissue extraction kits (Qiagen Inc.) originally proposed by Valentini et al. (2016) and further modified here (Supplementary methods S5.1).

Abundance estimates with qPCR

We assessed the concentration of Round Goby DNA using mtDNA and nuDNA markers in each eDNA sample using quantitative PCR (qPCR), which demonstrates better performance than digital droplet PCR (ddPCR) in Round Goby assays (Meredith Nevers, *pers. commun*). We quantified mtDNA using a species-specific primer set and fluorescent Taqman probe targeting a 147 bp region of the Round Goby cytochrome oxidase I (COI) gene developed in previous studies (Nathan, Jerde, Budny, & Mahon, 2015; Nevers et al., 2018). To quantify nuDNA, we used primers and probes targeting a 60 bp region of the *Nmel*505 microsatellite locus that were developed and validated in Chapter 4 of this dissertation (Table S4.2). All small-volume and large-volume eDNA samples and negative controls were assayed in 20 µL volumes: 10 µL TaqMan Environmental Master Mix 2.0 (ThermoFisher Scientific), 1.8 µL of each 10-µM primer (forward and reverse), 0.05 µL of 100-µM Taqman probe, 4.35 µL of PCR-grade water, and 2 µL of extracted DNA. We used the following thermocycling parameters: initial denaturation at 95°C for 10 min and 45 cycles of 95°C for 15 s and either 60°C (mtDNA) or 58°C (nuDNA) for 1 min. To quantify the amount of mtDNA and nuDNA copies/µL in eDNA samples, we generated standard curves from 10-fold serial dilutions (10–10^5 copies/µL) of synthetic gBlock Gene Fragments (Integrated DNA Technologies) with sequences that span the entire target gene region. We amplified the standard curve on each qPCR plate along with the eDNA sample extracts and used the Design and Analysis software (v2.6.0) to determine amplification efficiency and R^2 for each assay. All qPCR assays were

completed on the *QuantStudio 7 Pro* at the Cornell Institute of Biotechnology's Genomics Facility, with each sample amplified in triplicate reactions.

Due to poor amplification of nuDNA in both small-volume and large-volume eDNA samples, we suspected inhibition and diluted all samples ten-fold prior to amplification. After controlling for dilution, we did not observe a significant change in nuDNA copy numbers in diluted samples (paired t-test, $p = 0.43$). We therefore present the results of the undiluted samples here. The concentration of mtDNA and nuDNA in each sample was calculated as the average of the three amplification replicates (Ellison, English, Burns, & Keer, 2006).

Abundance estimates with genetic diversity

To quantify the amount of Round Goby genetic diversity in each sample, we sequenced a region of the mitochondrial genome and obtained a set of observed haplotypes per sample, which was then compared to population haplotype frequencies to estimate the number of individuals in each mixture of DNA (Sethi et al., 2019). Simulations in Chapter 2 of this dissertation demonstrated that large mixtures of individuals can be resolved using a single mitochondrial marker if enough genetic variation is present and can be detected (Figure 2.3). To maximize the amount of genetic variation in the target region, we designed a marker targeting a large region of the hypervariable d-loop control region, which contains elevated variation in comparison to other regions of the mitogenome (Gutnik, Walser, & Adrian-Kalchhauser, 2019). We aligned nine Round Goby mitochondrial sequences containing this gene region from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and designed primers to amplify a 565 bp region of the d-loop region using Geneious Prime (2020.0.5). The resulting

primers are GobyDLOOP-F: 5'-TGATTGTTGATGAATGAATGGCT-3' and GobyDLOOP-R: 5'-GGAGATTTTAACTCCCACCCC-3'.

We amplified the Round Goby d-loop region in each eDNA sample in triplicate 20 µL PCRs containing 0.8 µL of each 10-µM primer (forward and reverse), 4 µL of One*Taq* 5x buffer (New England Biolabs, Inc.), 0.4 µL of 10 mM dNTPs, 0.1 µL of One*Taq* DNA polymerase, 2 µL bovine serum albumin (BSA; 4 µg/µL), 9.9 µL of molecular H_2O, and 2 µL of eDNA sample. The cycling program for PCR included: initial denaturation at 95°C for 30 s and 45 cycles of 94°C for 30 s, 58°C for 45 s, and 68°C for 60 s, and final extension at 68ºC for 5 min. We visualized PCR products with electrophoresis on a 1% agarose gel stained with ethidium bromide. We then combined 5 µL of PCR product from triplicate reactions and uniquely barcoded each sample using Illumina Nextera XT tags. This second-stage PCR included 1.5 µL PCR product, 4 µL of One*Taq* 5x buffer (New England Biolabs, Inc.), 0.4 µL of 10 mM dNTPs, 0.1 µL of One*Taq* DNA polymerase, 12.4 µL of molecular H_2O, and 0.8 µL of each Index 1 (i7) and Index 2 (i5) adapters with a cycling program of six cycles of 95°C for 30 s, 62°C for 60 s, and 68°C for 60 s, with a final extension step of 68°C for 10 min. We pooled equimolar volumes of each sample based on the DNA concentration obtained from a Qubit 2.0 fluorimeter (Thermo Fisher Scientific) and performed size selection using Agencourt AMPure XP beads (reaction ratio AMPure beads 0.7: PCR product 1; Beckman Coulter Genomics). We submitted the final library to Cornell University's Institute of Biotechnology Genomics Facility where additional size selection was performed using Blue Pippin (Sage Science) followed by sequencing with the MiSeq Micro 500 bp kit (paired-end 2 × 250 bp).

Bioinformatic processing of Round Goby d-loop sequences in eDNA samples included trimming sequencing adapters with Trimmomatic v0.39 (Bolger, Lohse, & Usadel, 2014) and assessing the quality of sequences with FastQC version 0.11.8 (Andrews, 2010). Because sequence quality dropped off near the end of the sequences, we trimmed all F and R sequences to 225 bp. We then denoised the sequences and generated amplicon sequence variants (ASVs) using the divisive amplicon denoising algorithm, DADA2 (Callahan et al., 2016), which discarded sequences shorter than 100 bp or with > 2 expected errors, concatenated forward and reverse reads, and removed putative chimeras. Although we observed no contamination in field, laboratory, or sequencing blanks, we only retained sequences that were present at >1% frequency in a sample to reduce the potential for false positives. The final eDNA dataset contains information on haplotype read counts for each small-volume and large-volume eDNA sample.

Because genetic diversity-based abundance estimates are calculated by comparing the number of genetic variants identified in an eDNA sample to population-level frequencies, we estimated population haplotype frequencies using tissues from Round Gobies collected in Lake Huron from Chapter 4 of this dissertation (Table 4.1). Rare alleles in the specified population allele frequencies are important for estimating large numbers of individuals from eDNA (Chapter 2, Figure 2.3), so we also obtained Round Goby tissues from sites nearest to Lake Huron. In all, we sequenced the mitochondrial d-loop for 92 Round Gobies collected from Lake Huron (n = 30), Lake Michigan (n = 50), and Lake St. Clair (n = 12; Figure S5.2). The d-loop region was amplified in Round Goby tissues using the same primers and PCR conditions as described above, with the number of PCR cycles reduced to 25. All PCR

products were purified with Agencourt AMPure XP beads (Beckman Coulter Genomic) and Sanger sequenced at Cornell University's Institute of Biotechnology Genomics Facility.

We observed poor quality Sanger sequences with multiple out-of-sync ripples and overlapping peaks in electropherograms of several samples, which can result from high levels of heteroplasmy, and/or PCR slippage associated with the highly repetitive nature of the amplified fragment containing several microsatellite-type repeats. Because a consensus sequence could not be obtained for each individual with Sanger sequencing, we barcoded each sample with Illumina Nextera XT tags as described above and submitted tissues for sequencing on the Illumina platform with the MiSeq Nano 500 bp kit (paired-end 2 × 250 bp). Bioinformatic processing of tissue samples included the same steps as described for the eDNA samples, where the number of reads of each haplotype was determined for each individual. Following (Just, Irwin, & Parson, 2015), we determined the presence of heteroplasmy in instances where the frequency of the minor haplotype(s) accounted for > 10% of the reads of the major haplotype. We then estimated population frequencies based on haplotypes observed across all individuals, including instances of heteroplasmy.

The number of contributors to each eDNA sample was then determined using the DNA mixture model presented in Andres et al. (2021). This model uses combinatorics to calculate the likelihood that a putative number of contributors produces a set of observed haplotypes given the frequency of those haplotypes in the population. We applied this model to the haplotypes observed in each eDNA sample with population haplotype frequencies estimated from tissue sequences and calculated the maximum likelihood number of contributors from 1–50 individuals. This maximum likelihood estimate represents the number of Round Goby individuals whose DNA was captured in an eDNA sample.

Data analysis

Statistical analyses for this study were performed using R v4.1.2 (R Core Team, 2021). We compared the effect of AUV-based abundance estimates on qPCR concentration of both marker types and sample volumes as well as the number of genetic contributors in each small-volume and large-volume eDNA sample using linear mixed-effects models with the 'lmer' function in the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015). For small-volume eDNA-based abundance estimates, we specified sample and site specified as nested random effects. Because large-volume samples were not replicated at each site, we only specified site as a random effect in models of large-volume eDNA samples. We assessed the fit of the models based on the marginal R^2 values, which represents the variance explained by fixed effects in mixed-effect models, using the `MuMIn` package (Barton & Barton, 2015). We applied a log transformation to qPCR values to normalize model residuals.

Results

AUV images

We obtained usable AUV images at 10 of the 15 sites sampled for Round Gobies, with 5 sites exhibiting poor image quality primarily due to overexposure. At the remaining 10 sites, we obtained images at depths ranging from 0.5–40 m and observed fish in an average of 245.2 (s.d. = 142.9) out of 800 analyzed images per site, totaling 2,452 identified Round Gobies (Table 5.1). Round Goby density based on AUV images was negatively correlated with latitude ($r = -0.85$, $p = 0.002$), with no Round Gobies observed at the northernmost site and the maximum number of gobies observed at the southernmost site (Figure 5.1).

Association between goby density and DNA concentration

No Round Goby DNA was found in any of the negative controls using qPCR of the mitochondrial COI marker and standard curve efficiency ranged from 88–103% with R^2 ranging from 0.973–0.997. Round Goby mtDNA was detected at every sampled site, including the one site where no gobies were detected with AUV images (Figure 5.2). The concentration of mtDNA was higher per µL of extracted DNA in large-volume samples, with an average of 2,438 ± 2,689 copies/µL in large-volume eDNA samples and 271 ± 228 copies/µL in small-volume samples. However, after accounting for sample volume, small-volume (0.5 L) samples contained significantly higher DNA concentration per sampled mL than large-volume (100 L) samples (type III ANOVA with Satterthwaite's method; $F = 276.62$, $p < 0.001$), indicating a saturating effect of filter volume. The amount of mtDNA in eDNA was positively associated with goby abundance assessed with AUV images in small-volume samples ($F = 10.853$, adjusted $R^2 = 0.482$, $p = 0.020$; Figure 5.2B) but no association was found in large-volume samples ($F = 0.278$, adjusted $R^2 = 0.027$, $p = 0.61$; Figure 5.2D).

We did not amplify nuDNA in most eDNA samples, with an average of 2.04 (s.d. = 1.07) copy numbers of the *Nmel*505 microsatellite marker detected in just 4 of the 50 eDNA samples. All samples where nuDNA was detected were large-volume eDNA samples. With extremely low concentrations or non-detections of nuDNA across all eDNA samples, we did not assess the association of nuDNA with Round Goby abundance.

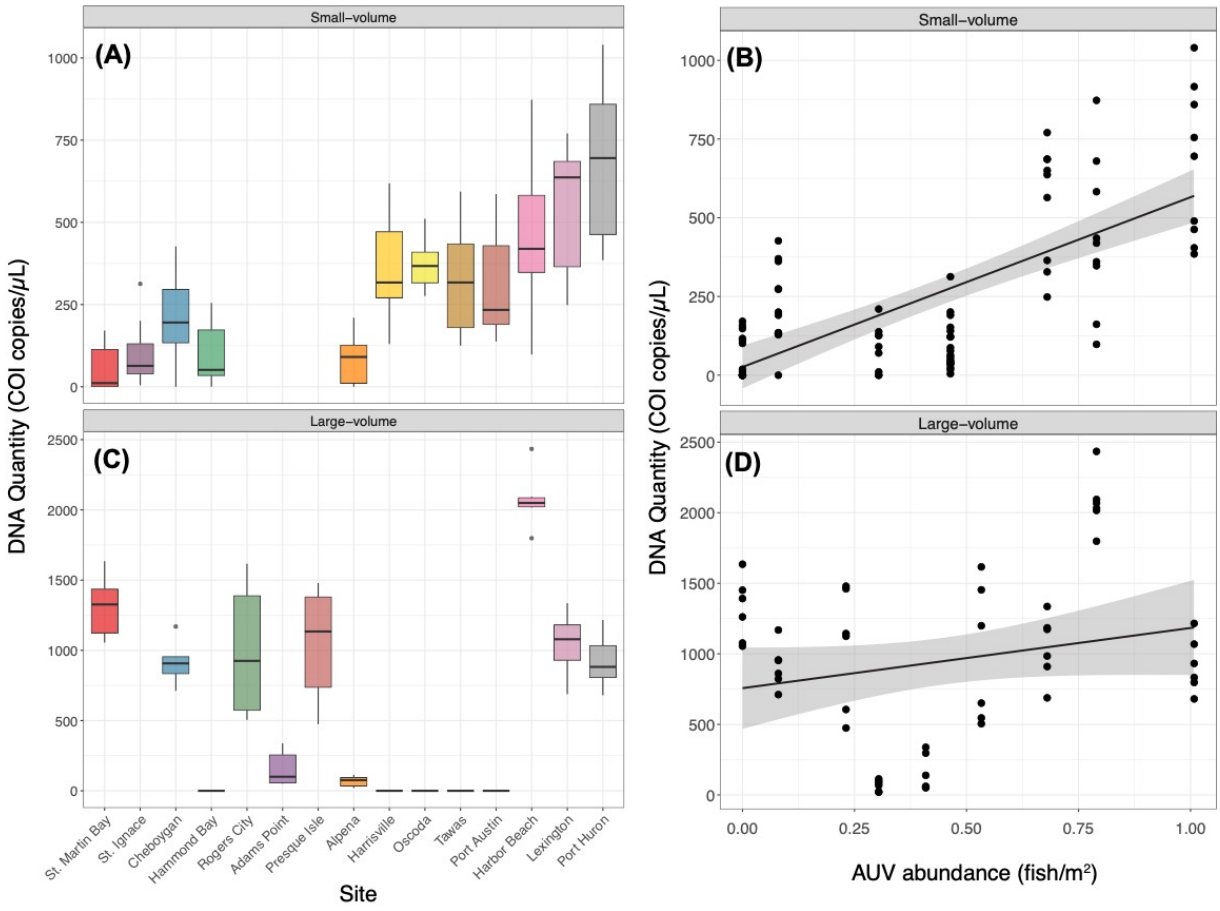

Figure 5.2. Boxplots show the quantity of Round Goby DNA per site estimated based on three qPCR amplification replicates of a mitochondrial COI marker using triplicate small-volume samples (A) and single large-volume samples (C). Median COI quantity values per site are represented by lines inside the boxes, first and third quartiles are represented by the top and bottom of the boxes, and 1.5 * the interquartile range is shown with upper and lower whiskers). The relationship between Round Goby abundance estimated from AUV images (fish/m^2) and mtDNA quantity (COI copies/µL extracted sample) is shown for (B) small-volume eDNA samples and (D) large-volume eDNA samples. Black lines were fitted based on the coefficients of the linear mixed effects model and shaded areas represent the 95% confidence interval.

Challenges with quantifying genetic diversity

We processed a total of 475,404 reads from 92 Round Goby tissue samples. Genotyping individuals by sequencing revealed high levels of heteroplasmy in this gene region of Round Gobies, with 62 individuals exhibiting at least two (and up to four) distinct d-loop sequences at a read frequency exceeding the 10% threshold. Including instances of heteroplasmy, we observed 34 haplotypes across all individuals, 21 of which were shared by at least one individual. Major haplotypes were shared across all sampled sites (Figure S5.2).

For eDNA samples, we processed a total of 4,772,174 demultiplexed sequence reads, but only 2,780 of these reads were determined to be Round Goby d-loop sequences. To establish the cause of low sequencing success, we tracked reads through the DADA2 pipeline and discovered that the vast majority of sequence reads were removed due to issues of sequence length and quality (Figure S5.3). We also inspected taxonomic assignments for all sequences using the BLASTn algorithm (Camacho et al., 2009), nucleotide and taxonomy (nt/taxdb) databases from NCBI (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2005; Federhen, 2012), and MEGAN v.6.23.3 (Huson et al., 2016), and discovered that many sequences passing quality filters can be traced to bacterial origin, indicating issues of primer specificity (Figure S5.3).

Due to these complications with sequence quality and specificity, we only amplified Round Goby d-loop haplotypes at 4 of the 12 sites surveyed with small-volume eDNA samples. In the samples where goby d-loop sequences were observed, we estimated an average of 2.5 (s.d. = 1.3, range = 1–5) Round Gobies per sample using the contributor estimation from the DNA mixture model (Figure 5.3). No large-volume eDNA samples

contained any Round Goby d-loop sequences, likely due to the quality and specificity problems mentioned above combined with additional inhibition issues. Because so few samples exhibited positive amplification of d-loop haplotypes, we did not assess the association between AUV-based and genetic diversity-based estimates of Round Goby abundance.

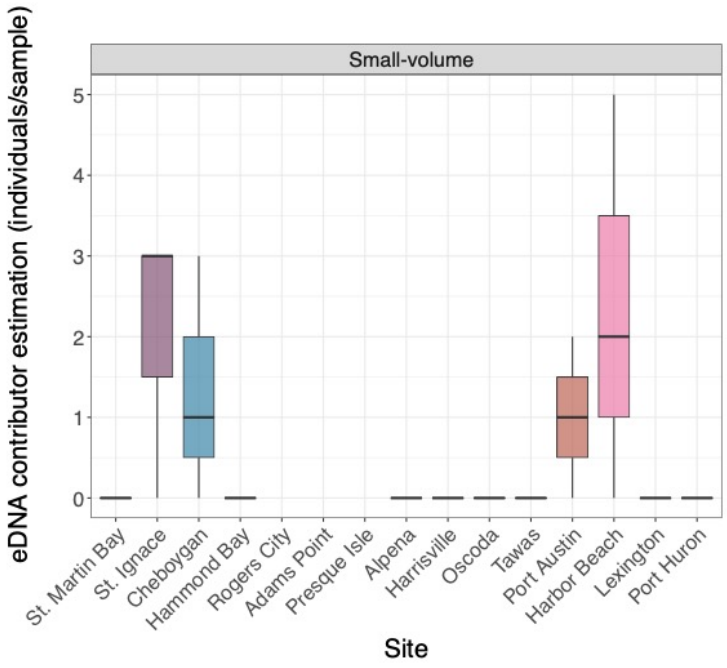

Figure 5.3. Boxplots showing the number of Round Gobies estimated to have contributed to small-volume eDNA samples per sites estimated based on the amount of genetic variation in the mitochondrial d-loop detected in eDNA. Median values per site are shown by lines inside the boxes, first and third quartiles are represented by the top and bottom of the boxes, and 1.5 * the interquartile range is shown with upper and lower whiskers). Large-volume eDNA samples did not successfully amplify Round Goby d-loop haplotypes and are not shown.

Discussion

Preventing the spread and minimizing the undesirable impacts of invasive species requires efficient monitoring of population abundance across broad spatial and temporal scales. With this research, we test the performance of eDNA-based estimates of population abundance to investigate use of eDNA as a tool for predicting species-specific abundance in natural systems. Previous research investigating the relationship between the concentration of eDNA in water samples and species abundance have found generally positive correlations, but due to the many processes impacting the amount of DNA detected in environmental samples, much of the variation remains unexplained (average $R^2 = 0.50$ in natural environments; Yates et al., 2019). We attempted to account for some of this variation by eliminating the influence of within- and among-individual differences in DNA production by targeting nuclear genetic markers and estimating abundance with genetic diversity. In comparison to the more common approach of estimating abundance using the concentration of mitochondrial DNA, we anticipated these approaches would exhibit greater similarity to Round Goby estimates derived from benthic images. Although low DNA concentrations and problems with marker specificity prevented a robust assessment of these novel eDNA-based approaches, we found significant similarity between Round Goby abundance from AUV images and quantity of mtDNA, providing support for the use of eDNA approaches based on mtDNA as a useful tool in population assessment, monitoring, and aquatic conservation.

Because different cell types contain different numbers of mitochondria but only a single nucleus, we expected that the concentration of nuDNA would exhibit lower variance and closer similarity to Round Goby abundance than mtDNA. However, we were unable to

amplify nuDNA out of most of the samples, likely as a consequence of low DNA concentrations. We amplified both mtDNA and nuDNA in Chapter 4 of this dissertation (Figure 4.1) and found that the ratio of mtDNA:nuDNA in natural settings is almost 200:1. However, eDNA samples in that study were collected in nearshore environments in shallow depths and 2 L of water was filtered per sample. In contrast, we filtered 0.5 mL of water for small-volume samples and collected surface water samples in depths of approximately 5 m. Accordingly, mtDNA copy numbers were much lower in samples analyzed in this study (average 271 copies/μL) than those in Chapter 4 (average 1,300 copies/μL). Assuming the ratio of mtDNA to nuDNA holds across different environments, nuDNA concentrations in this study were below the detectable limit of quantification with qPCR. Although qPCR is a sensitive approach for detecting species DNA at moderate–high concentrations, ddPCR demonstrates significantly better detection at low concentrations (Doi et al., 2015; Jerde & Mahon, 2015; Jerde et al., 2016). We therefore recommend researchers consider marker detectability and assay sensitivity when designing studies to target nuclear genetic markers.

If nuDNA is the desired target, it may be more feasible to amplify multi-copy nuclear genes from environmental samples rather than the single-copy markers used in this study. Previous research has demonstrated that multi-copy nuclear ribosomal RNA markers such as ITS1 have similar or higher concentrations and detectability in eDNA than mitochondrial markers (Jo, Arimoto, Murakami, Masuda, & Minamoto, 2020; Jo, Takao, & Minamoto, 2022; Minamoto et al., 2017). Such markers may have the benefit of reduced sensitivity to different cell types in environmental samples while being sufficiently abundant to be reliably detected in eDNA samples. More research on the exploration of different target gene regions,

particularly nuclear markers, will be important to better understand the optimal genetic markers for assessing species abundance.

To test the impact of sample volume on the relationship between eDNA estimates and species abundance, we collected large-volume and small-volume eDNA samples at each site. Although we anticipated that large-volume samples would be less sensitive to sampling stochasticity and exhibit a closer relationship with Round Goby abundance than small-volume sampling, we found the opposite to be true. Small-volume sampling exhibited a significantly positive correlation between mtDNA concentration and AUV-based estimates of Round Goby abundance, while no relationship was found in large-volume sampling. Interestingly, we also did not observe amplification of nuDNA in most of the large-volume eDNA samples (100 L), even though mtDNA concentrations in these samples (average 2,438 copies/μL) were much higher than in the eDNA samples in Chapter 4 where nuDNA was reliably amplified. We measured no Round Goby mtDNA in the four large-volume samples that contained high levels of sediments (Table 5.1), even though eDNA has been shown to degrade more slowly and exhibit higher concentrations in sediments than in surface waters (Turner, Uy, & Everhart, 2015). We therefore suspect that large-volume samples may be experiencing inhibition due to increased sediment loads in large volumes of filtered water. Dissolved substances and sediments in the water column can bind DNA and inhibit PCR, influencing eDNA detection rates (Albers, Jensen, Bælum, & Jacobsen, 2013; Stoeckle et al., 2017). Although we did not directly test for inhibition, we used reagents that reduce the impacts of inhibition (Jane et al., 2015) and diluted extracted DNA 1:10 (McKee, Spear, & Pierson, 2015). However, further control for inhibition is possible (Coyne et al., 2005) and may be required for large-volume sampling when inhibiting substances are present. Other studies that

have had success filtering large volumes of water of up to 100 L (Cilleros et al., 2019; Valentini et al., 2016) have been primarily conducted in lotic waters that may contain fewer inhibiting substances than the sites sampled in this study.

We also experienced challenges in amplifying the mitochondrial d-loop from eDNA samples that would allow for a test of the potential to use genetic variation to estimate the absolute abundance of Round Gobies captured in an eDNA sample. The d-loop is part of the mitochondrial control region, which is the only major non-coding region of the mitochondrial genome in most species and is the site of replication origin (Clayton, 1982). In many species, this region exhibits highly variable DNA sequences and length polymorphisms, often with regions of tandem repetition are associated with high levels of heteroplasmy (up to 100% of individuals exhibiting more than one form of mtDNA; Arnason & Rand, 1992; Buroker et al., 1990; Zhang & Hewitt, 1997). We observed heteroplasmy arising from both sequence and length variations in 67% of the sequenced individuals, with seven individuals exhibiting four distinct d-loop sequences with frequencies greater than 10%. The presence of such high levels of sequence and length heteroplasmy greatly complicate the interpretation of DNA mixtures, which typically assume predictable ploidy at a given genetic marker (Melton, Holland, & Holland, 2012). Furthermore, true heteroplasmy can be challenging to distinguish from other causes of apparent intraindividual sequence variation such as sequencing errors and non-specific priming (Just et al., 2015). Therefore, although the mitochondrial control region is a common target for eDNA studies assessing intraspecific genetic variation (Baker, Steel, Nieukirk, & Klinck, 2018; Sigsgaard et al., 2016) and contains high levels of variation that make it suitable for assessments of organismal abundance (Yoshitake et al., 2021; Yoshitake

et al., 2019), estimating the number of individuals in a sample using DNA mixtures frameworks may be limited if heteroplasmy is prevalent.

On other hand, other regions of the mitochondrial genome may not contain sufficient variation to estimate more than a few individuals. Other genetic markers, such as nuclear microsatellites, may also contain very high levels of variation. However, as discussed above and in Chapter 4, nuclear genetic markers may be challenging to amplify from environmental samples, and low detections of genetic variants would negatively bias estimate of the number of genetic contributors collected in a sample. Reliably estimating genetic variation in eDNA samples has only recently been explored and remains challenging (Sigsgaard et al., 2020), and more research is needed to determine the best markers for detecting genetic variation for use in abundance estimation frameworks.

In targeting different genetic markers and estimating genetic variation in eDNA samples, we attempted to reduce the amount of variation in eDNA-based estimates of species abundance that arise from contributions of different amounts of DNA from different individuals. However, even if this variation is removed, other source of variation will be present that may influence the amount and detectability of eDNA in a sample. Careful consideration of the physical conditions and mechanisms influencing eDNA shedding, decay, transport, resuspension, and settling is required to obtain robust estimates of organism abundance (Barnes & Turner, 2015; Barnes et al., 2014; Nevers et al., 2018; Shogren et al., 2017). Ultimately, in order for approaches using the abundance of eDNA to progress from relative to absolute species abundance and density estimates, a thorough understanding of DNA dynamics in the sampled environment may be required.

All abundance estimates, whether derived from conventional sampling with nets or advanced survey methods such as eDNA and AUV images, have their own benefits as well as sources of bias that may vary depending on the conditions and characteristics of the sampled site (Pierce et al., 2001). For instance, although benthic images provide more accurate abundance estimates of Round Gobies than trawling (Johnson, Allen, et al., 2005), AUV images may not be able to distinguish Round Gobies from deepwater sculpins (*Myoxocephalus thompsonii*) or slimy sculpins (*Cottus cognatus*) due to low resolution or turbid environments. While this is not be major issue in this study, as Round Gobies and are spatially separated from other benthic species in invaded habitats (Pothoven, 2018), this may be a challenge in the future as Round Gobies spread into deeper habitats, increasing their habitat overlap with sculpin species (Volkel et al., 2021).

Taxonomic resolution can also be a problem when estimating abundance with eDNA, as non-specific primers can lead to amplification of DNA from non-target species, artificially inflating species abundance estimates. Species specificity may not be a major concern for markers with widespread usage and expansive reference sequences such as COI (Porter & Hajibabaei, 2018); however, the development of new genetic markers will require careful consideration of primer specificity. Importantly, neither approach used in this study involves physical capture of the species. Although this may be preferable in some cases, for instance when assessing the status of threatened or endangered species, many studies require physical collection of individuals for age, diet, or detailed genetic analysis. Because different survey methods provide different information with different efficiency and biases, multiple methods may be combined to minimize effort while maximizing the amount of information obtained

(Andres, Lambert, Lodge, Andrés, & Jackson, 2022; Underwood, Rosen, Engås, & Eriksen, 2014).

Estimating species abundance with noninvasive methods is a challenging but useful approach for collecting biological data that can be used to efficiently evaluate ecological status or track the efficacy of management actions. Given the ongoing threat of introduced and invasive species to biodiversity, monitoring invasive species presence and abundance with efficient approaches such as eDNA is essential for minimizing their spread and impacts. Advanced survey methods such as AUV assessments are already being used in Great Lakes environmental research (Austin, 2013; Bennion et al., 2019), with growing potential to sample environments at greater resolution with reduced effort. Environmental DNA methods, while still under development in many systems, also show potential as a rapid and low-cost alternative or supplement to existing species surveys. Although we experienced challenges amplifying nuclear genetic markers and detecting d-loop variation, we found an association between the concentration of DNA and Round Goby abundance evaluated with AUV images. These results provide support for the utility of eDNA in quantifying species abundance using DNA concentrations and highlight the importance of considering species specificity and DNA concentration when assessing species abundance using alternative genetic markers.

Acknowledgements

Funding for this project was provided by the University of Michigan Cooperative Institute for Great Lakes Research (CIGLR), through the National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement NA17OAR4320152, the National Science Foundation (NSF) Coastal SEES program (grant number 1748389), and the Department of Defense (DoD) Strategic Environmental Research and Development Program (SERDP; grant number RC19-1004). We thank Maria Haileselassie, Diana Tapia, George Rogalskyj, Miranda Andres, and Joe Boschert for assistance with collecting specimens in the field. All animals were collected and euthanized according to the Cornell IACUC Animal Care and Use Procedure (ACUP 306.02) and local collecting permits.

References

Aday, D. D. (2007). The presence of and invasive macrophyte (*Phragmites australis*) does not influence juvenile fish habitat use in a freshwater estuary.

Adolfson, J., & Berghage, T. (1974). *Perception and performance under water*: John Wiley & Sons.

Albers, C. N., Jensen, A., Bælum, J., & Jacobsen, C. S. (2013). Inhibition of DNA polymerases used in Q-PCR by structurally different soil-derived humic substances. *Geomicrobiology Journal, 30*(8), 675-681.

Andres, K. J., Lambert, T. D., Lodge, D. M., Andrés, J., & Jackson, J. R. (2022). Combining sampling gear to optimally inventory species highlights the efficiency of eDNA metabarcoding. *In press, Environmental DNA*.

Andres, K. J., Sethi, S. A., Duskey, E., Lepak, J. M., Rice, A. N., Estabrook, B. J., . . . Scofield, A. E. (2020). Seasonal habitat use indicates that depth may mediate the potential for invasive Round Goby impacts in inland lakes. *Freshwater Biology, 65*(8), 1337-1347.

Andres, K. J., Sethi, S. A., Lodge, D. M., & Andrés, J. (2021). Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Molecular Ecology, 30*(3), 685-697.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: *Babraham Bioinformatics*, Babraham Institute, Cambridge, United Kingdom.

Arim, M., Abades, S. R., Neill, P. E., Lima, M., & Marquet, P. A. (2006). Spread dynamics of invasive species. *Proceedings of the National Academy of Sciences, 103*(2), 374-378.

Arnason, E., & Rand, D. (1992). Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, Gadus morhua. *Genetics, 132*(1), 211-220.

Austin, J. (2013). The potential for autonomous underwater gliders in large lake research. *Journal of Great Lakes Research, 39*, 8-13.

Baker, C. S., Steel, D., Nieukirk, S., & Klinck, H. (2018). Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. *Frontiers in Marine Science, 5*, 133.

Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics, 17*(1), 1-17. doi:10.1007/s10592-015-0775-4

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M. (2014). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology, 48*(3), 1819-1827.

Barton, K., & Barton, M. K. (2015). Package 'mumin'. *Version, 1*(18), 439.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Bennion, D. H., Warner, D. M., Esselman, P. C., Hobson, B., & Kieft, B. (2019). A comparison of chlorophyll a values obtained from an autonomous underwater vehicle to satellite-based measures for Lake Michigan. *Journal of Great Lakes Research, 45*(4), 726-734.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research, 33*(suppl_1), D34-D38.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Bradley, B. A., Allen, J. M., O'Neill, M. W., Wallace, R. D., Bargeron, C. T., Richburg, J. A., & Stinson, K. (2018). Invasive species risk assessments need more consistent spatial abundance data. *Ecosphere, 9*(7), e02302.

Brown, J. H. (1984). On the relationship between abundance and distribution of species. *American Naturalist*, 255-279.

Brown, R. W., Ebener, M., & Gorenflo, T. (1999). Great Lakes commercial fisheries: historical overview and prognosis for the future. *Great Lakes fisheries policy and management: a binational perspective. Michigan State University Press, East Lansing*, 307-354.

Buroker, N., Brown, J., Gilbert, T., O'hara, P., Beckenbach, A., Thomas, W., & Smith, M. (1990). Length heteroplasmy of sturgeon mitochondrial DNA: an illegitimate elongation model. *Genetics, 124*(1), 157-163.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics, 10*(1), 1-9.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., . . . Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources, 19*(1), 27-46.

Clayton, D. A. (1982). Replication of animal mitochondrial DNA. *Cell, 28*(4), 693-705.

Coyne, K. J., Handy, S. M., Demir, E., Whereat, E. B., Hutchins, D. A., Portune, K. J., . . . Cary, S. C. (2005). Improved quantitative real-time PCR assays for enumeration of harmful algal species in field samples using an exogenous DNA reference standard. *Limnology and Oceanography: Methods, 3*(9), 381-391.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., . . . Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology, 26*(21), 5872-5895.

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T., & Minamoto, T. (2017). Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology, 62*(1), 30-39.

Doi, H., Takahara, T., Minamoto, T., Matsuhashi, S., Uchii, K., & Yamanaka, H. (2015). Droplet digital polymerase chain reaction (PCR) outperforms real-time PCR in the detection of environmental DNA from an invasive fish species. *Environmental Science & Technology, 49*(9), 5601-5608.

Douglas, M. M., & O'Connor, R. A. (2003). Effects of the exotic macrophyte, para grass (*Urochloa mutica*), on benthic and epiphytic macroinvertebrates of a tropical floodplain. *Freshwater Biology, 48*(6), 962-971.

Elgersma, K. J., & Ehrenfeld, J. G. (2011). Linear and non-linear impacts of a non-native plant invasion on soil microbial community structure and function. *Biological Invasions, 13*(3), 757-768.

Ellison, S. L., English, C. A., Burns, M. J., & Keer, J. T. (2006). Routes to improving the reliability of low level DNA analysis using real-time PCR. *BMC biotechnology, 6*(1), 1-11.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Research, 40*(D1), D136-D143.

Freedman, J. A., Fuller, P., Benson, A., Maynard, E., Neilson, M. E., Larson, J., . . . Sturtevant, R. (2022, 4/26/2022). *Neogobius melanostomus* (Pallas, 1814): U.S. Geological Survey, Nonindigenous Aquatic Species Database. Retrieved from https://nas.er.usgs.gov/queries/factsheet.aspx?SpeciesID=713

Gaston, K. J., & Lawton, J. H. (1990). Effects of scale and habitat on the relationship between regional distribution and local abundance. *Oikos*, 329-335.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., . . . Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution, 7*(11), 1299-1307.

Gutnik, S., Walser, J.-C., & Adrian-Kalchhauser, I. (2019). Long-read sequencing of benthophilinae mitochondrial genomes reveals the origins of Round Goby mitogenome re-arrangements. *Mitochondrial DNA Part B, 4*(1), 408-412.

Hondorp, D. W., O'Brien, T. P., Esselman, P., & Roseman, E. F. (2022). *Status and trends of the Lake Huron prey fish community, 1976-2019*. Great Lakes Fishery Commission.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., . . . Tappu, R. (2016). MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology, 12*(6), e1004957.

Jackson, M. C., Ruiz-Navarro, A., & Britton, J. R. (2015). Population density modifies the ecological impacts of invasive species. *Oikos, 124*(7), 880-887.

Jane, S. F., Wilcox, T. M., McKelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., . . . Whiteley, A. R. (2015). Distance, flow and PCR inhibition: e DNA dynamics in two headwater streams. *Molecular Ecology Resources, 15*(1), 216-227.

Janssen, J., & Jude, D. J. (2001). Recruitment Failure of Mottled Sculpin *Cottus bairdi* in Calumet Harbor, Southern Lake Michigan, Induced by the Newly Introduced Round Goby *Neogobius melanostomus*. *Journal of Great Lakes Research, 27*(3), 319-328. doi:10.1016/s0380-1330(01)70647-8

Jerde, C. L., & Mahon, A. R. (2015). Improving confidence in environmental DNA species detection. *Molecular Ecology Resources*, *15*(3), 461-463.

Jerde, C. L., Olds, B. P., Shogren, A. J., Andruszkiewicz, E. A., Mahon, A. R., Bolster, D., & Tank, J. L. (2016). Influence of stream bottom substrate on retention and transport of vertebrate environmental DNA. *Environmental Science & Technology, 50*(16), 8770-8779.

Jo, T., Arimoto, M., Murakami, H., Masuda, R., & Minamoto, T. (2020). Estimating shedding and decay rates of environmental nuclear DNA with relation to water temperature and biomass. *Environmental DNA, 2*(2), 140-151.

Jo, T., Takao, K., & Minamoto, T. (2022). Linking the state of environmental DNA to its application for biomonitoring and stock assessment: Targeting mitochondrial/nuclear genes, and different DNA fragment lengths and particle sizes. *Environmental DNA, 4*(2), 271-283.

Johnson, T. B., Allen, M., Corkum, L. D., & Lee, V. A. (2005). Comparison of methods needed to estimate population size of Round Gobies (*Neogobius melanostomus*) in western Lake Erie. *Journal of Great Lakes Research, 31*(1), 78-86. doi:10.1016/s0380-1330(05)70239-2

Johnson, T. B., Bunnell, D. B., & Knight, C. T. (2005). A potential new energy pathway in central Lake Erie: the Round Goby connection. *Journal of Great Lakes Research, 31*, 238-251.

Jude, D. J., Reider, R. H., & Smith, G. R. (1992). Establishment of Gobiidae in the Great Lakes Basin. *Canadian Journal of Fisheries and Aquatic Sciences, 49*(2), 416-421. doi:10.1139/f92-047

Just, R. S., Irwin, J. A., & Parson, W. (2015). Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Science International: Genetics, 18*, 131-139.

King, R., Ray, J., & Stanford, K. (2006). Gorging on gobies: beneficial effects of alien prey on a threatened vertebrate. *Canadian Journal of Zoology, 84*(1), 108-115.

Kolar, C. S., & Lodge, D. M. (2001). Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution, 16*(4), 199-204.

Krakowiak, P. J., & Pennuto, C. M. (2008). Fish and macroinvertebrate communities in tributary streams of eastern Lake Erie with and without round gobies (*Neogobius melanostomus*, Pallas 1814). *Journal of Great Lakes Research, 34*(4), 675-689.

Lauer, T. E., Allen, P. J., & McComish, T. S. (2004). Changes in Mottled Sculpin and Johnny Darter Trawl Catches after the Appearance of Round Gobies in the Indiana Waters of

Lake Michigan. *Transactions of the American Fisheries Society, 133*(1), 185-189. doi:10.1577/t02-123

Leino, J. R., & Mensinger, A. F. (2016). The benthic fish assemblage of the soft-bottom community of the Duluth-Superior Harbor before and after round goby invasion (1989–2011). *Journal of Great Lakes Research, 42*(4), 829-836.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

McElroy, M. E., Dressler, T. L., Titcomb, G. C., Wilson, E. A., Deiner, K., Dudley, T. L., . . . Jerde, C. L. (2020). Calibrating environmental DNA metabarcoding to conventional surveys for measuring fish species richness. *Frontiers in Ecology and Evolution, 8*, 276.

McKee, A. M., Spear, S. F., & Pierson, T. W. (2015). The effect of dilution and the use of a post-extraction nucleic acid purification column on the accuracy, precision, and inhibition of environmental DNA samples. *Biological Conservation, 183*, 70-76.

Melton, T., Holland, C., & Holland, M. (2012). Forensic mitochondria dna analysis: current practice and future potential. *Forensic Science Review, 24*(2), 101.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp *Cyprinus carpio*. *Molecular Ecology Resources, 17*(2), 324-333.

Nathan, L. R., Jerde, C. L., Budny, M. L., & Mahon, A. R. (2015). The use of environmental DNA in invasive species surveillance of the Great Lakes commercial bait trade. *Conservation Biology, 29*(2), 430-439.

Nevers, M. B., Byappanahalli, M. N., Morris, C. C., Shively, D., Przybyla-Kelly, K., Spoljaric, A. M., . . . Roseman, E. F. (2018). Environmental DNA (eDNA): A tool for quantifying the abundant but elusive Round Goby (*Neogobius melanostomus*). *PLoS One, 13*(1), e0191720. doi:10.1371/journal.pone.0191720

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics, 26*(3), 419-420.

Pierce, C. L., Corcoran, A. M., Gronbach, A. N., Hsia, S., Mullarkey, B. J., & Schwartzhoff, A. J. (2001). Influence of diel period on electrofishing and beach seining assessments of littoral fish assemblages. *North American Journal of Fisheries Management, 21*(4), 918-926.

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., & Waits, L. P. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences, 70*(8), 1123-1130.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics, 52*(3), 273-288.

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One, 13*(9), e0200177.

Pothoven, S. A. (2018). Seasonal feeding ecology of co-existing native and invasive benthic fish along a nearshore to offshore gradient in Lake Michigan. *Environmental Biology of Fishes, 101*(7), 1161-1174.

Pothoven, S. A., & Madenjian, C. P. (2013). Increased piscivory by lake whitefish in Lake Huron. *North American Journal of Fisheries Management, 33*(6), 1194-1202.

R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rao, Y. R., & Schwab, D. J. (2007). Transport and mixing between the coastal and offshore waters in the Great Lakes: a review. *Journal of Great Lakes Research, 33*(1), 202-218.

Ricciardi, A., & MacIsaac, H. J. (2000). Recent mass invasion of the North American Great Lakes by Ponto–Caspian species. *Trends in Ecology & Evolution, 15*(2), 62-65.

Riley, S. C., Roseman, E. F., Nichols, S. J., O'Brien, T. P., Kiley, C. S., & Schaeffer, J. S. (2008). Deepwater Demersal Fish Community Collapse in Lake Huron. *Transactions of the American Fisheries Society, 137*(6), 1879-1890. doi:10.1577/t07-141.1

Robin, E. D., & Wong, R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *Journal of Cellular Physiology, 136*(3), 507-513.

Sala, O. E., Stuart Chapin, F., Armesto, J. J., Berlow, E., Bloomfield, J., Dirzo, R., . . . Wall, D. H. (2000). Global biodiversity scenarios for the year 2100. *Science, 287*(5459), 1770-1774.

Schaeffer, J. S., Bowen, A., Thomas, M., French III, J. R., & Curtis, G. L. (2005). Invasion history, proliferation, and offshore diet of the round goby *Neogobius melanostomus* in western Lake Huron, USA. *Journal of Great Lakes Research, 31*(4), 414-425.

Sethi, S. A., Larson, W., Turnquist, K., Isermann, D., & Kembel, S. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution, 10*(1), 109-119. doi:10.1111/2041-210x.13079

Shogren, A. J., Tank, J. L., Andruszkiewicz, E., Olds, B., Mahon, A. R., Jerde, C. L., & Bolster, D. (2017). Controls on eDNA movement in streams: Transport, Retention, and Resuspension. *Scientific Reports, 7*(1), 5065. doi:10.1038/s41598-017-05223-1

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Moller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA-Current status and future perspectives. *Evolutionary Applications, 13*(2), 245-262. doi:10.1111/eva.12882

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., . . . Thomsen, P. F. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution, 1*(1), 4. doi:10.1038/s41559-016-0004

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S. T., . . . Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution, 8*(5), 635-645. doi:10.1111/2041-210x.12683

Steingraeber, M., Runstrom, A., & Thiel, P. (1996). *Round Goby (Neogobius melanostomus) distribution in the Illinois Waterway System of metropolitan Chicago*: US Fish and Wildlife Service, La Crosse Fishery Resources Office.

Steinhart, G. B., Marschall, E. A., & Stein, R. A. (2004). Round Goby predation on smallmouth bass offspring in nests during simulated catch-and-release angling. *Transactions of the American Fisheries Society, 133*(1), 121-131.

Stockwell, J. D., Yule, D. L., Gorman, O. T., Isaac, E. J., & Moore, S. A. (2006). Evaluation of bottom trawls as compared to acoustics to assess adult lake herring (*Coregonus artedi*) abundance in Lake Superior. *Journal of Great Lakes Research, 32*(2), 280-292.

Stoeckle, B. C., Beggel, S., Cerwenka, A. F., Motivans, E., Kuehn, R., & Geist, J. (2017). A systematic approach to evaluate the influence of environmental conditions on eDNA detection success in aquatic ecosystems. *PLoS One, 12*(12), e0189119.

Strickler, K. M., Fremier, A. K., & Goldberg, C. S. (2015). Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation, 183*, 85-92.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*: Oxford University Press.

Thomas, A. C., Howard, J., Nguyen, P. L., Seimon, T. A., & Goldberg, C. S. (2018). eDNA Sampler: A fully integrated environmental DNA sampling system. *Methods in Ecology and Evolution, 9*(6), 1379-1385.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation, 183*, 4-18. doi:10.1016/j.biocon.2014.11.019

Troth, C. R., Sweet, M. J., Nightingale, J., & Burian, A. (2021). Seasonality, DNA degradation and spatial heterogeneity as drivers of eDNA detection dynamics. *Science of the Total Environment, 768*, 144466.

Turner, C. R., Uy, K. L., & Everhart, R. C. (2015). Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation, 183*, 93-102.

Underwood, M. J., Rosen, S., Engås, A., & Eriksen, E. (2014). Deep vision: An in-trawl stereo camera makes a step forward in monitoring the pelagic community. *PLoS One, 9*(11), e112304.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., . . . Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology, 25*(4), 929-942. doi:10.1111/mec.13428

Volkel, S. L., Robinson, K. F., Bunnell, D. B., Connerton, M. J., Holden, J. P., Hondorp, D. W., & Weidel, B. C. (2021). Slimy sculpin depth shifts and habitat squeeze following the round goby invasion in the Laurentian Great Lakes. *Journal of Great Lakes Research, 47*(6), 1793-1803.

Yates, M. C., Fraser, D. J., & Derry, A. M. (2019). Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. *Environmental DNA, 1*(1), 5-13.

Yoshitake, K., Fujiwara, A., Matsuura, A., Sekino, M., Yasuike, M., Nakamura, Y., . . . Watabe, S. (2021). Estimation of tuna population by the improved analytical pipeline of unique molecular identifier-assisted HaCeD-Seq (haplotype count from eDNA). *Scientific Reports, 11*(1), 1-12.

Yoshitake, K., Yoshinaga, T., Tanaka, C., Mizusawa, N., Reza, M., Tsujimoto, A., . . . Watabe, S. (2019). HaCeD-Seq: a novel method for reliable and easy estimation about the fish population using haplotype count from eDNA. *Marine Biotechnology, 21*(6), 813-820.

Zhang, D.-X., & Hewitt, G. M. (1997). Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology, 25*(2), 99-120.

APPENDIX 1

Supplementary Information for Chapter 1:

Combining sampling gear to optimally inventory species highlights the efficiency of eDNA metabarcoding

Table S1.1. Occurrences of all detected taxa among sampling approaches.

Table S1.2. Breakdown of estimated effort (person-hours of work) required to obtain one sample for each sampling gear type.

Table S1.3. Breakdown of estimated cost ($USD) required to obtain one sample for each sampling gear type.

Table S1.4. Breakdown of in-house eDNA cost estimates.

Figure S1.1 Heat map showing species' detection rates (in purple) for each sampling method.

Figure S1.2. Correlations between species occurrence.

Figure S1.3. Site-specific species richness.

Table S1.1. Occurrences of all detected taxa among sampling approaches (eDNA metabarcoding, electrofishing, fyke netting, gillnetting, and seining) and historical datasets (since 1990).

Family	Scientific name	Common name	Historical	eDNA	Electro-fishing	Fyke netting	Gill-netting	Seining
Centrarchidae	*Ambloplites rupestris*	Rock Bass	X	X	X	X	X	X
Ictaluridae	*Ameiurus nebulosus*	Brown Bullhead	X	X	X	X	X	X
Sciaenidae	*Aplodinotus grunniens*	Freshwater Drum	X	X	X	X	X	X
Catostomidae	*Catostomus commersonii*	White Sucker	X	X	X	X	X	X
Clupeidae	*Dorosoma cepedianum*	American Gizzard Shad	X	X	X	X	X	X
Esocidae	*Esox niger*	Chain Pickerel	X	X	X	X	X	X
Centrarchidae	*Lepomis gibbosus*	Pumpkinseed	X	X	X	X	X	X
Centrarchidae	*Micropterus dolomieu*	Smallmouth Bass	X	X	X	X	X	X
Moronidae	*Morone americana*	White Perch	X	X	X	X	X	X
Gobiidae	*Neogobius melanostomus*	Round Goby	X	X	X	X	X	X
Percidae	*Perca flavescens*	Yellow Perch	X	X	X	X	X	X
Percidae	*Sander vitreus*	Walleye	X	X	X	X	X	X
Fundulidae	*Fundulus diaphanus*	Banded Killifish	X	X	X	X	–	X
Centrarchidae	*Lepomis cyanellus*	Green Sunfish	X	X	X	X	–	X
Centrarchidae	*Lepomis macrochirus*	Bluegill	X	X	X	X	–	X
Centrarchidae	*Micropterus salmoides*	Largemouth Bass	X	X	X	X	–	X
Cyprinidae	*Notemigonus crysoleucas*	Golden Shiner	X	X	X	X	–	X
Cyprinidae	*Notropis atherinoides*	Emerald Shiner	X	X	X	X	–	X
Cyprinidae	*Notrropis hudsonius*	Spottail Shiner	X	X	X	X	–	X
Percidae	*Percina caprodes*	Logperch	X	X	X	X	–	X
Ictaluridae	*Ameiurus natalis*	Yellow Bullhead	X	X	X	X	–	–
Amiidae	*Amia calva*	Bowfin	X	X	X	X	–	–
Catostomidae	*Erimyzon oblongus*	Creek Chubsucker	X	X	X	X	–	–
Lotidae	*Lota lota*	Burbot	X	X	X	X	–	–
Catostomidae	*Moxostoma valenciennesi*	Greater Redhorse	X	X	X	–	X	–
Cyprinidae	*Cyprinus carpio*	Common Carp	X	X	X	–	–	X
Atherinopsidae	*Labidesthes sicculus*	Brook Silverside	X	X	X	–	–	X
Percopsidae	*Percopsis omiscomaycus*	Trout Perch	X	X	X	–	–	–
Acpenseridae	*Acipenser fulvescens*	Lake Sturgeon	X	X	–	–	X	–
Ictaluridae	*Ictalurus punctatus*	Channel Catfish	X	X	–	–	X	–
Clupeidae	*Alosa pseudoharengus*	Alewife	X	X	–	–	–	–
Gasterosteidae	*Culaea inconstans*	Brook Stickleback	X	X	–	–	–	–
Esocidae	*Esox lucius*	Northern Pike	X	X	–	–	–	–
Percidae	*Etheostoma flabellare*	Fantail Darter	X	X	–	–	–	–
Percidae	*Etheostoma olmstedi*	Tessellated Darter	X	X	–	–	–	–
Catostomidae	*Hypentelium nigricans*	Northern Hogsucker	X	X	–	–	–	–
Cyprinidae	*Notropis heterolepis*	Blacknose Shiner	X	X	–	–	–	–
Cyprinidae	*Rhinichthys atratulus*	Blacknose Dace	X	X	–	–	–	–
Cyprinidae	*Rhinichthys cataractae*	Longnose Dace	X	X	–	–	–	–
Cyprinidae	*Semotilus atromaculatus*	Creek Chub	X	X	–	–	–	–
Umbridae	*Umbra limi*	Central Mudminnow	X	X	–	–	–	–
Centrarchidae	*Pomoxis nigromaculatus*	Black Crappie	X	–	X	X	–	–
Lepisosteidae	*Lepisosteus osseus*	Longnose Gar	X	–	X	–	–	X
Cyprinidae	*Pimephales promelas*	Fathead Minnow	X	–	X	–	–	X
Cyprinidae	*Cyprinella analostana*	Satinfin Shiner	X	–	–	X	–	X

Cyprinidae	*Pimephales notatus*	Bluntnose Minnow	X	–	–	X	–	X
Cyprinidae	*Hybognathus hankinsoni*	Brassy Minnow	X	–	–	–	–	X
Clupeidae	*Alosa aestivalis*	Blueback Herring	X	–	–	–	–	–
Anguillidae	*Anguilla rostrata*	American Eel	X	–	–	–	–	–
Cyprinidae	*Campostoma anomalum*	Central Stoneroller	X	–	–	–	–	–
Cyprinidae	*Carassius auratus*	Goldfish	X	–	–	–	–	–
Catostomidae	*Catostomus catostomus*	Longnose Sucker	X	–	–	–	–	–
Cyprinidae	*Clinostomus elongatus*	Redside Dace	X	–	–	–	–	–
Salmonidae	*Coregonus artedi*	Cisco	X	–	–	–	–	–
Cottidae	*Cottus bairdi*	Mottled Sculpin	X	–	–	–	–	–
Cottidae	*Cottus cognatus*	Slimy Sculpin	X	–	–	–	–	–
Cyprinidae	*Cyprinella spiloptera*	Spotfin Shiner	X	–	–	–	–	–
Esocidae	*Esox lucius x Esox masquinongy*	Tiger Musky	X	–	–	–	–	–
Cyprinidae	*Exoglossum maxillingua*	Cutlips Minnow	X	–	–	–	–	–
Cyprinidae	*Hybognathus regius*	Eastern Silvery Minnow	X	–	–	–	–	–
Cyprinidae	*Luxilus cornutus*	Common Shiner	X	–	–	–	–	–
Cyprinidae	*Margariscus margarita*	Pearl Dace	X	–	–	–	–	–
Moronidae	*Morone chrysops*	White Bass	X	–	–	–	–	–
Catostomidae	*Moxostoma macrolepidotum*	Shorthead Redhorse	X	–	–	–	–	–
Cyprinidae	*Nocomis biguttatus*	Hornyhead Chub	X	–	–	–	–	–
Cyprinidae	*Notropis rubellus*	Rosyface Shiner	X	–	–	–	–	–
Cyprinidae	*Notropis stramineus*	Sand Shiner	X	–	–	–	–	–
Ictaluridae	*Noturus flavus*	Stonecat	X	–	–	–	–	–
Ictaluridae	*Noturus insignis*	Margined Madtom	X	–	–	–	–	–
Salmonidae	*Oncorhynchus mykiss*	Rainbow Trout	X	–	–	–	–	–
Percidae	*Percina maculata*	Blackside Darter	X	–	–	–	–	–
Petromyzontidae	*Petromyzon marinus*	Sea Lamprey	X	–	–	–	–	–
Salmonidae	*Salmo salar*	Atlantic Salmon	X	–	–	–	–	–
Salmonidae	*Salmo trutta*	Brown Trout	X	–	–	–	–	–
Salmonidae	*Salvelinus fontinalis*	Brook Trout	X	–	–	–	–	–
Cyprinidae	*Scardinius erythrophthalmus*	Rudd	X	–	–	–	–	–
Cyprinidae	*Semotilus corpralis*	Fallfish	X	–	–	–	–	–

Table S1.2. Breakdown of estimated effort (person-hours of work) required to obtain one sample for each sampling gear type.

eDNA	Number of people	Effort (hours)	Total effort (hours)	Effort per site (hours)	Notes
Field collection	2	6	12	0.5	Including travel time
Sample filtration	2	6	12	0.5	
Lab work	1	20	20	0.8	10 hours DNA extraction + 10 hours library preparation
Data analysis	1	6	6	0.2	
Total			**50**	**2.0**	
Electrofishing	**Number of people**	**Effort (hours)**	**Effort per site (hours)**		**Notes**
Electrofishing run	2	0.5	1		Two 15-minute transect passes
Identify fish on shore	2	0.5	1		
Travel time	2	1	2		Avg. travel time to and from sites
Total			**4**		
Fyke netting	**Number of people**	**Effort (hours)**	**Effort per site (hours)**		**Notes**
Deploy nets	2	0.5	1		
Retrieve nets	2	0.5	1		
Identify fish on shore	2	0.5	1		
Travel time	2	0.3	0.6		Avg. travel time to and from sites/3 sites per day
Total			**3.6**		
Gill netting	**Number of people**	**Effort (hours)**	**Effort per site (hours)**		**Notes**
Deploy net	2	0.5	1		
Retrieve net	2	0.5	1		
Remove fish from net	6	1	6		
Identify fish on shore	6	0.5	3		
Travel time	2	1	2		Avg. travel time to and from sites
Total			**13**		
Seining	**Number of people**	**Effort (hours)**	**Effort per site (hours)**		**Notes**
Deploy net	2	0.5	1		
Identify fish	2	0.5	1		
Travel time	2	0.25	0.5		Avg. travel time to and from sites (seine sites are very near to dock)
Total			**2.5**		

Table S1.3. Breakdown of estimated cost ($USD) required to obtain one sample for each sampling gear type.

eDNA	Cost (USD)	Cost per site (USD)	Notes
In-house			
Labor	1000	40	Effort * $20/hour
Boat usage fee	60	2.4	Covers maintenance, fuel and depreciation; $60 daily boat fee, 25 sites/day
Cost per sample (in-house)	2739	110	In-house eDNA cost estimates based on Bálint et al. (2018); see next table
Total (in-house)		**152**	
External			
Labor	480	19	12 hours sample collection, 12 hours filtration
Boat usage fee	60	2.4	Covers maintenance, fuel and depreciation; $60 daily boat fee, 25 sites/day
Cost per sample (external)	--	262	External eDNA cost based on https://www.smith-root.com/edna/laboratory-services
Total (external)		**284**	
Electrofishing	**Cost per site (USD)**		**Notes**
Boat usage fee	200		Covers maintenance, fuel and depreciation; $200 daily boat fee, 1 site/day
Labor	80		Effort * $20/hour
Total	**280**		
Fyke netting	**Cost per site (USD)**		**Notes**
Boat usage fee	20		Covers maintenance, fuel and depreciation; $60 daily boat fee, 3 sites/day
Labor	72		Effort * $20/hour
Total	**92**		
Gillnetting	**Cost per site (USD)**		**Notes**
Boat usage fee	60		Covers maintenance, fuel and depreciation; $60 daily boat fee, 1 site/day
Labor	260		Effort * $20/hour
Total	**320**		
Seining	**Cost per site (USD)**		**Notes**
Boat usage fee	20		Covers maintenance, fuel and depreciation; $60 daily boat fee, 3 sites/day
Labor	50		Effort * $20/hour
Total	**70**		

Table S1.4. Breakdown of in-house eDNA cost estimates based on Bálint et al. (2018).

Item	Cost ($USD)	Notes
Negative controls	4	
Samples extracted at once	25	
Extraction cost per reaction	2	
Extraction costs	**58**	
Extraction time per batch (hr)	10	
Labor cost (USD/hr)	20	
Extraction time cost	**200**	
Number of repliacate PCRs per sample	4	
PCR cost (USD/reaction)	2	
PCR costs	**232**	
PCR setup time per plate (hr)	2	
PCR setup costs	**40**	
Library preparation costs	**100**	
Sequencing cost	**1,869**	Cornell Core Facilities (BRC): NextSeq 500 Mid output kit (paired-end 2×150 bp)
Bioinformatic time (hr)	12	
Bioinformatic costs	**240**	
Total costs	**2739**	

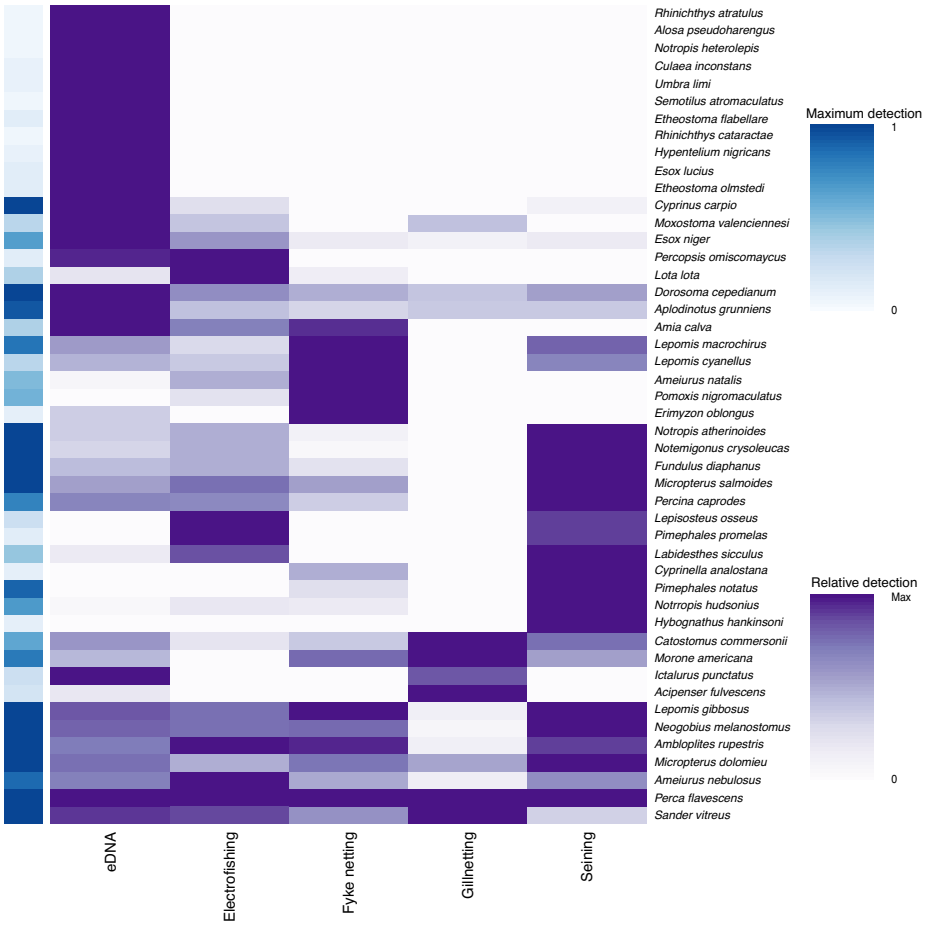

Figure S1.1 Heat map showing species' detection rates (in purple) for each of the five sampling methods: eDNA, electrofishing, fyke netting, gillnetting, and seining. Each species' detection rate is displayed relative to the gear with the highest fraction of samples in which the species was detected; this maximum detection fraction is shown at left in blue.

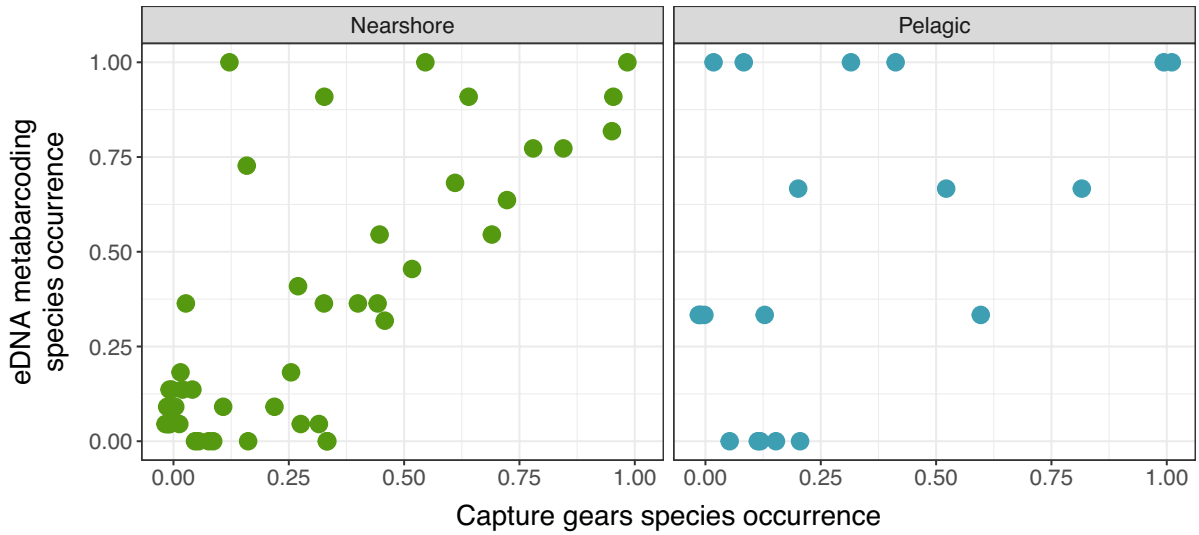

Figure S1.2. Correlations between species occurrence based on capture-based survey methods and eDNA metabarcoding in nearshore and pelagic habitats.

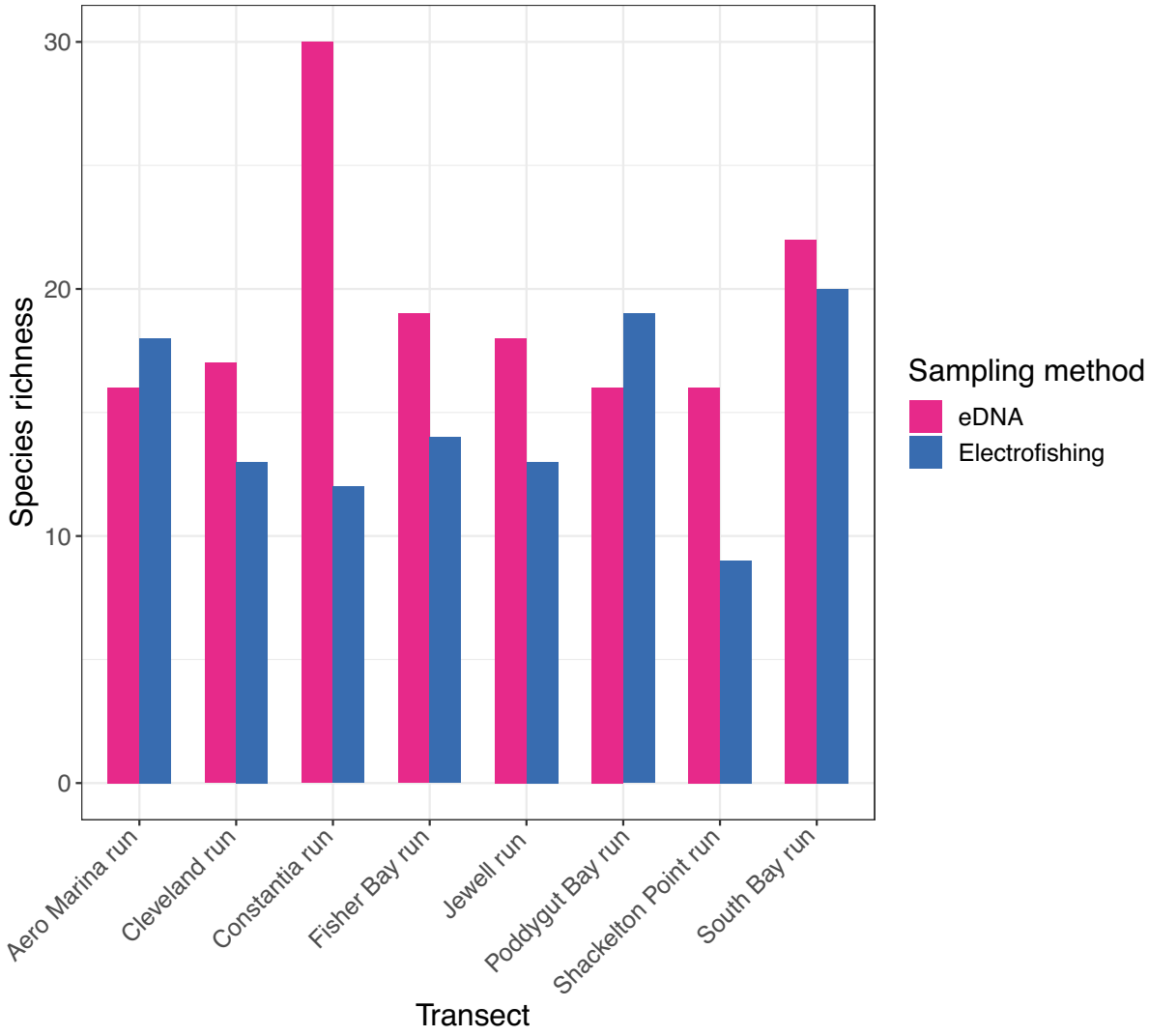

Figure S1.3. Site-specific species richness detected with eDNA metabarcoding and electrofishing sampling methods at each of eight electrofishing transect sites.

APPENDIX 2

Supplementary Information for Chapter 2:

Detecting and analyzing intraspecific genetic variation with environmental DNA

Supplementary methods S2.1. Mitochondrial genotyping methods.

Figure S2.1. Per-locus likelihood estimate.

Supplementary methods S2.1. Mitochondrial genotyping methods.

To sequence the *Elacatinus lori* mitogenome, we used PrimerSelect (DNASTAR) to design non-overlapping primer pairs across the entire mitochondrial genome. PCR fragments were generally ~450 bp and did not exceed 470 bp. We designed two multiplex mixes and placed mito-genomically adjacent PCR products into different multiplex groups. Multiplex PCR products were amplified with QIAGEN Multiplex PCR kits. Amplified products were pooled within samples and barcoded with Illumina Nextera dual barcodes. After barcoding, sample aliquots were pooled, cleaned with 0.7x Ampure XP, and diluted to a concentration of 2nM. We sequenced the pooled barcoded samples on an Illumina MiSeq (2 x 250 bp reads).

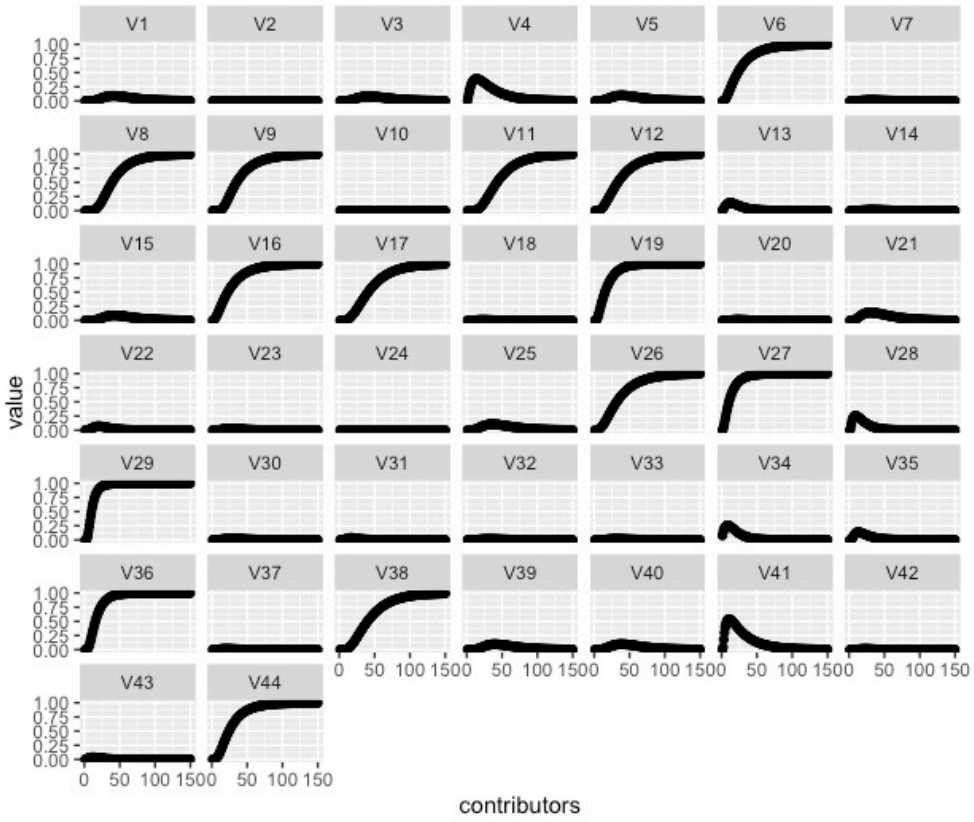

Figure S2.1. Per-locus likelihood estimate across 1:150 putative contributors in a mixture of DNA containing microsatellite genotypes from 40 individuals. The loci that fail (i.e., monotonically increase across all putative contributors and never reach a maximum likelihood) are removed when calculating the likelihood product across all loci for microsatellite markers.

APPENDIX 3

Supplementary Information for Chapter 3:

Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples

Table S3.1. Summary information for 35 autosomal microsatellite loci.

Figure S3.1. Sites of mesocosm/field collections of Round Gobies and BIC vs. number of clusters.

Figure S3.2. Correlation of allele read depth between duplicate eDNA samples per mesocosm.

Figure S3.3. Correlation of allele read frequencies between duplicate eDNA samples per mesocosm.

Figure S3.4. Mean read depth of field eDNA samples, mesocosm eDNA samples, and tissue samples.

Figure S3.5. Total per-sample and per-locus read depth for eDNA samples in each mesocosm density treatment.

Table S3.1. Summary information for 35 autosomal microsatellite loci used in the study, including the primer name, repeat motif, number of motif repeats, forward and reverse primer sequences, multiplex number, and allelic richness in genotyped Round Goby tissues. Denoted loci (†) were excluded from analysis due to poor amplification or deviation from Hardy-Weinberg equilibrium.

Primer	Tetramer repeat	Repeat count	Forward sequence (5'-3')	Reverse sequence (3'-5')	Multi-plex	Allelic richness
Nmel140	AGAT	11	ATGTGCTGGATTGTCTCATTGG	AGCTTGTTAATATGGCACGGTG	4	11
Nmel155†	ATCC	15	ACATAACTTGGTGCTCATTCTGG	GGTGCAATGTGATTTGGGCAG	4	3
Nmel185†	AGAT	12	ATGTGAGACTATTTCCTTCGGG	TTTCCTCCTCTGGCTTTGTGTG	4	2
Nmel248†	AGAT	13	ACTGAATGACAACTGAATTTCCC	TCATGTTAAATTCAGATCCAGGTC	1	21
Nmel262	ATCC	10	CGACAGCTGTAGTTCAACTCAC	AAACGGACAAGTGGAAGAAG	1	7
Nmel299	AGAT	11	TTATGCTCACTACTGCCACTTC	AGGAAACAAGGAAGGGACAAAC	1	12
Nmel32	AGAT	11	CTCCAAACCTTCTTACTGCTGC	AACCCTCACCTGAATCAAGCAC	4	4
Nmel344	AGAT	14	GCTATAATGTTGTGTGCAGTGC	ACACTGCCTTAATGATTCAAAGG	5	9
Nmel351	AGAT	16	TTTGGTCACTGGAATTACAATACG	CATATTGGCTTCAGTAGATGTTTG	1	12
Nmel361†	AAAG	12	CGAGGATCCTGGCAATTGGG	CAACTTTCGGTGCAAACATGAC	5	8
Nmel363	AAGT	10	TCAAAGAGACCTAAGCCAGTCC	CTGTTTGGCTCGGACAACATG	1	9
Nmel403	AGAT	11	GTCTTGCTGCATAAACCAAAGC	GGACCGATCTGTATATAGGAGCC	5	10
Nmel405	AGAT	18	TCACAGCTAAGACCAACAACC	AATGGACTAACAGTTCGCTAC	5	15
Nmel411	ATCC	15	GGAAACAAGGTCGCAGGTAATG	CAGCAACGGACAAGTGGAAG	2	8
Nmel422	ATCC	11	CTCCCATCTCCCATCCAAGTAC	TACAAAGCGCGATATCAGTTGG	2	3
Nmel505	AGAT	10	CCCACTTCAATGTTCTGCACTC	GCTTTCACCTCATTCATACAGC	5	5
Nmel549	AGAT	11	TGCATGATATAGGACCTTTAAAGC	TGGACTGTAAGCCATAATCTTC	6	11
Nmel625	ATCC	12	CAGACAAAGAGCGGTTCAGAAG	TGATGGTGGGTGAAGTACATGG	6	10
Nmel660	ATCC	12	CTGGACACAAGCTGCAAGTG	CTATGTGATTTGGGCAGCTCC	2	2
Nmel726†	AAAG	11	CTCTCAGCGTATACAGAGGCC	CAGGTCCACTAATTCGATGTCC	2	8
Nmel729	ACAG	13	ACAAGTGCATTAGTGTCATGGG	AAGGCCTAACAAACAGCAGATG	2	5
Nmel746	AAAG	17	GTTTATGCCCTGGACATCTGC	GCTCGAGTCTTAAATGCAGATTG	6	6
Nmel810	ATCC	12	AACACAGCGTCAAATCTCTCAG	ATTACTAGTAGGTGACGGGTCG	7	7
Nmel815	AGAT	15	GGGTGTATTGTCTAGTCTCTGG	CAGGCTCATGTTAAGGGTTCAG	6	12
Nmel821	ACAT	11	CCACCAATCATGTTACACAGGG	AGTTCAGAGGCAGCCAGTTG	7	7
Nmel89	ACAT	10	TCTACTTCAGTTGGCTATGATTC	TAAGAGCAGGCTTAAGACCCAC	7	6
Nmel914	AGAT	12	TTGTGGACAAGGGCTGAACAG	CTATACGCACAGCTTCCTCACC	3	6
Nmel990	AGAT	13	AAATGCTGTTATACTGAGGCGG	CTCAGGGCTCCAACTACTCC	3	12
Nmel994	AGAT	11	CTCTCTGACATGCTCCAAGGAC	TTCTCAATAACTTTATCTCGGACC	3	11
Nmel1103†	AAAG	13	GCCAAGTCCTATCTCGCAATG	CGCCCTCGGTACTGTTATAAAG	7	19
Nmel1132	AGAT	10	TCCTGGATGAACACTACAAGGC	AGGACGTTTCGCTTTGATCTTC	7	12
Nmel1462	AAAG	10	CGATACCCAATATAGCGGACGG	AGTCAGAAGGAAACACATGCAG	5	8
Nmel1486	ATCC	13	TTCCCATTGTAACAGCAGAGAG	AATGTGATTTGGGCAGCTCG	6	14
Nmel1531†	AAAC	10	TGTAACAGGAGCCAAATCAGATG	GCCCTGGGTACAAACTATCTTG	7	17
Nmel1566	AGAT	14	ATTTCGGATCCTCGTATCTGAC	AATGATATGGGAAATGCGAGCG	3	19

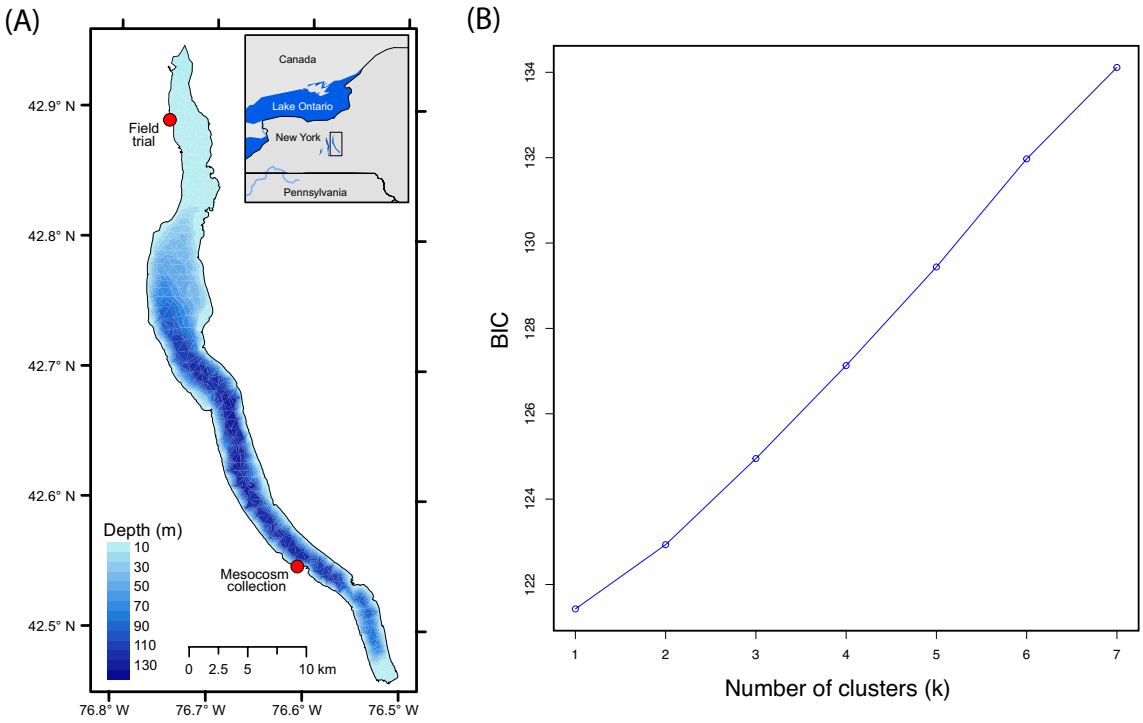

Figure S3.1. Sites of Round Goby (*Neogobius melanostomus*) sampling in Cayuga Lake, New York, USA for the mesocosm experiment and field trial. Environmental DNA samples from Cayuga Lake were collected at the field trial site. (B) We determined the Round Gobies collected from the two sites are panmictic, as the optimal number of clusters indicated by the lowest value of Bayesian Information Criterion (BIC) is 1.

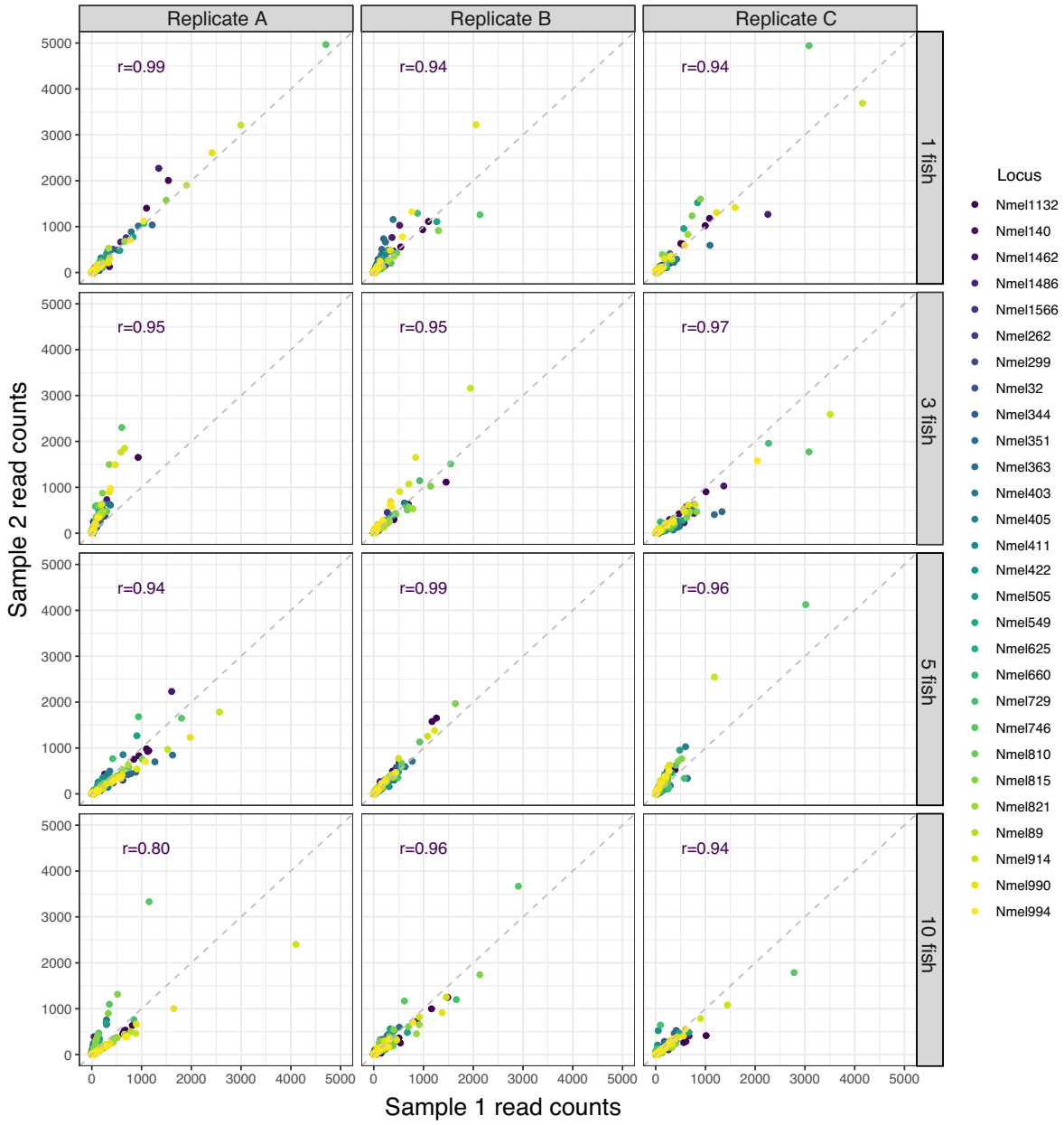

Figure S3.2. Correlation of allele read depth between duplicate eDNA samples per mesocosm. Mesocosms consist of 3 replicates (A, B, or C) of each density treatment (1, 3, 5, or 10 fish). Pearson's correlation coefficient (r) is indicated in each panel. Colors represent each of 28 loci. Diagonal lines represent a 1:1 relationship between eDNA sample 1 and sample 2 read counts.

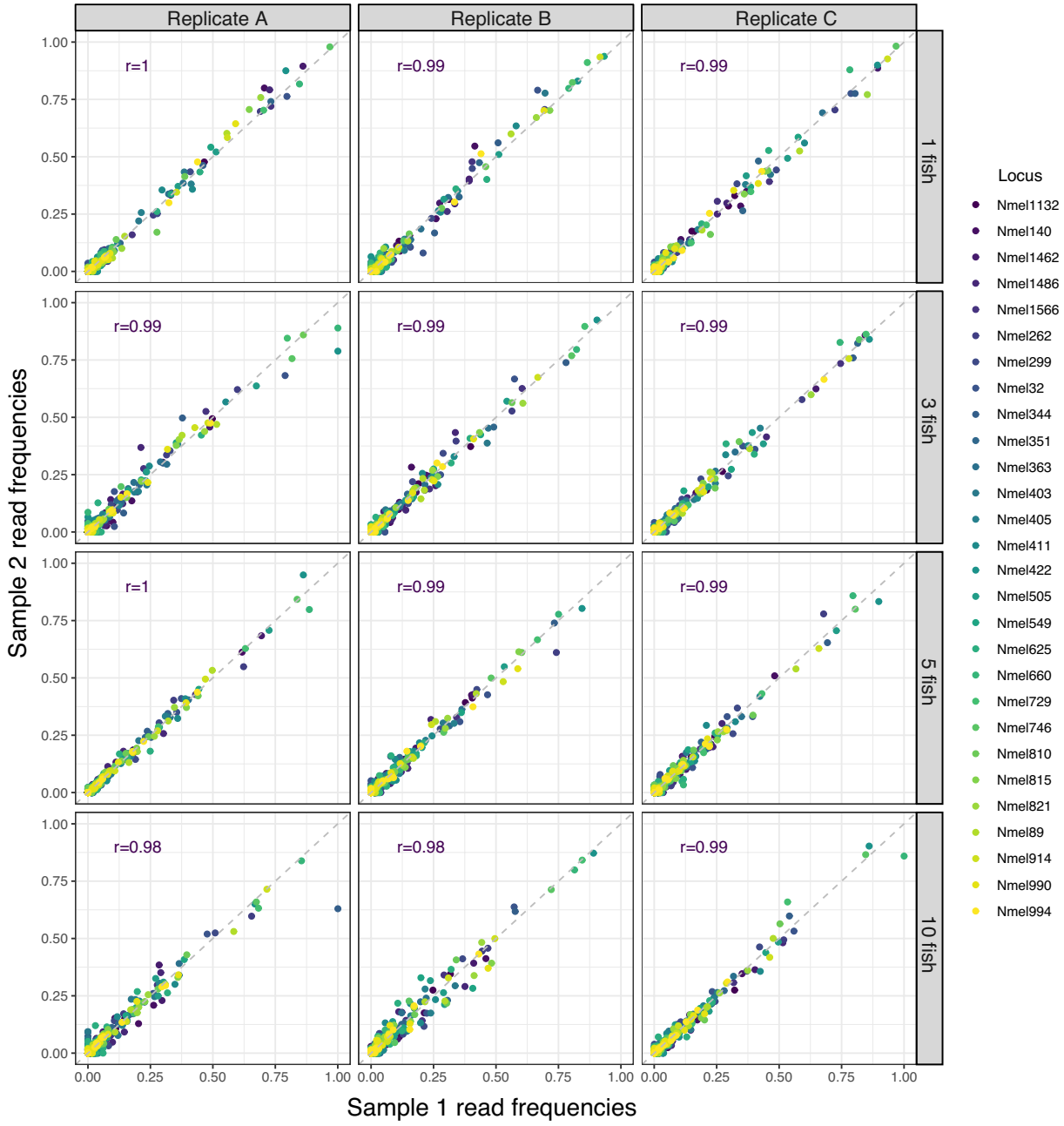

Figure S3.3. Correlation of allele read frequencies between duplicate eDNA samples per mesocosm. Mesocosms consist of 3 replicates (A, B, or C) of each density treatment (1, 3, 5, or 10 fish). Pearson's correlation coefficient (r) is indicated in each panel. Colors represent each of 28 loci. Diagonal lines represent a 1:1 relationship between eDNA sample 1 and sample 2 read frequencies.

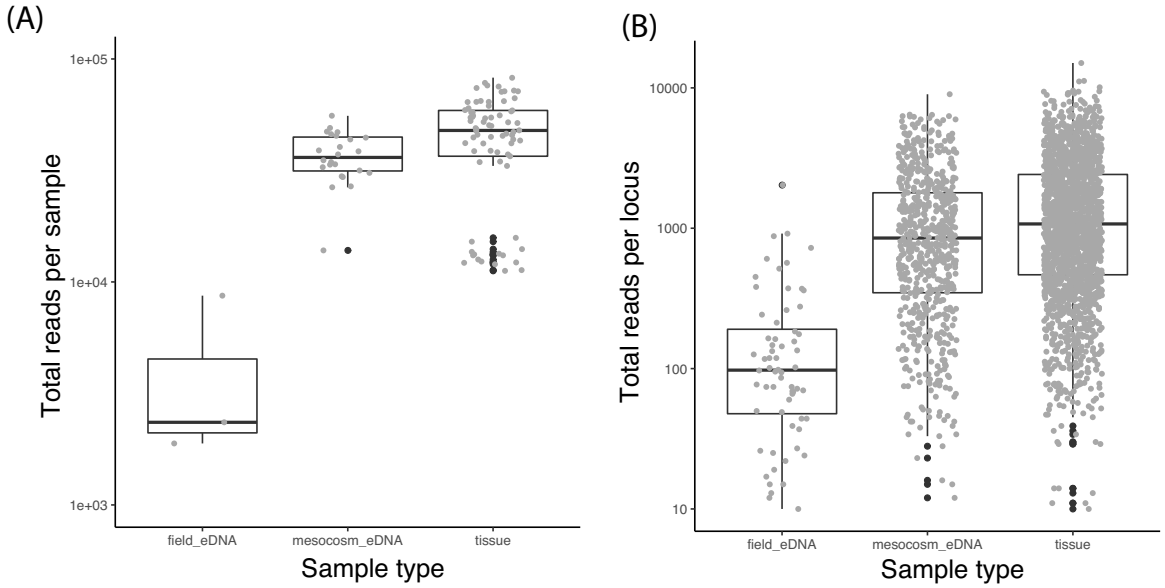

Figure S3.4. (A) Total reads per sample and (B) total reads per locus across all alleles in 28 loci. Sample types include the three replicate field eDNA samples, 24 mesocosm eDNA samples, and tissues sampled from 73 individuals in the experiment.

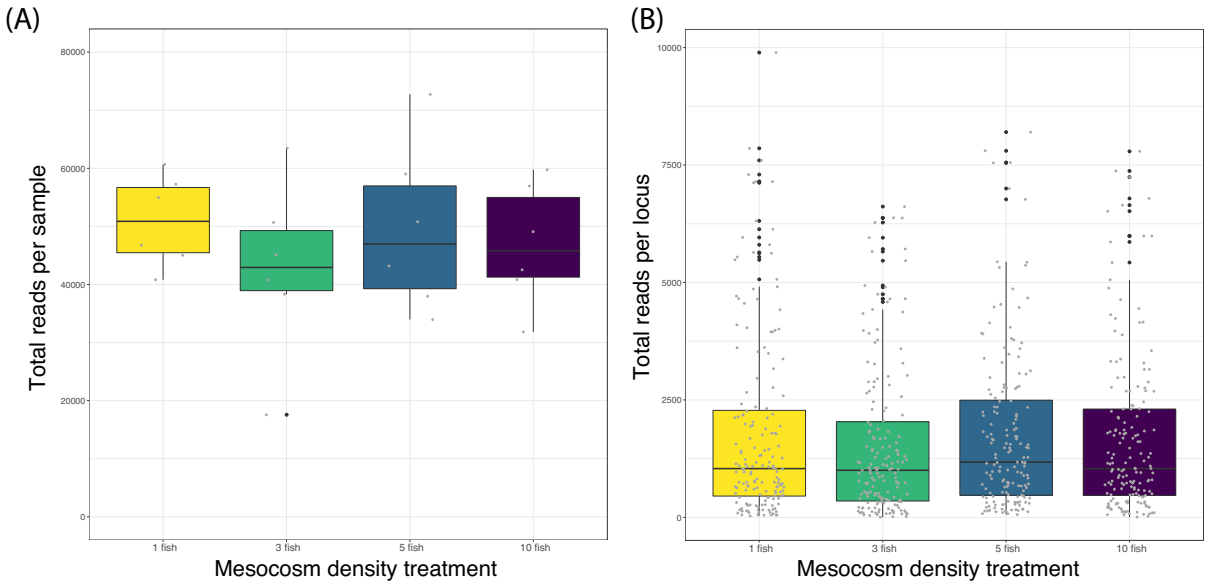

Figure S3.5. (A) Total reads per sample and (B) total reads per locus across all alleles in 28 loci for eDNA samples in each mesocosm density treatment (1, 3, 5, or 10 fish).

APPENDIX 4

Supplementary Information for Chapter 4:

Environmental DNA detects population genetic diversity and structure

with microsatellite markers

Table S4.1. Overview of eDNA and tissue samples collected in this study.

Table S4.2. The qPCR primer/probe assays used in this study.

Table S4.3. Summary information for 28 microsatellite loci used in the study.

Figure S4.1. Boxplot showing the mean ratio of mtDNA to nuDNA across eDNA samples per population.

Figure S4.2. Pairwise F_{ST} comparisons based on tissue-based genetic analysis 27 microsatellite loci for Round Goby populations in North America.

Figure S4.3. Spatial analysis of Round Goby population structure based on Round Goby tissues.

Figure S4.4. Read depth, number of alleles, and number of loci detected in each eDNA sample. Triplicate samples are colored per site.

Figure S4.5. Accumulation curves showing the number of microsatellite alleles expected in a given number of Round Goby tissue samples and eDNA samples.

Figure S4.6. Summary statistics for population-level genetic data collected from genotyped Round Goby tissues and eDNA samples.

Figure S4.7. Correlation of allele frequencies per site.

Table S4.1. Overview of eDNA and tissue samples collected in this study. Table includes sampling sites, collection dates, total number of microsatellite sequences obtained from each eDNA sample, number of alleles recovered from each eDNA sample, and results from qPCR analysis of the average estimated eDNA concentration across technical replicates.

Environmental DNA										
Date	**Body of water**	**Type**	**Sample_ID**	**Site_ID**	**Latitude**	**Longitude**	**mtDNA quantity**	**nuDNA quantity**	**Reads**	**Retained**
7/25/19	Seneca Lake	Blank	SENBL	SEN	42.869	-76.941	0.0	0.0	3	–
7/25/19	Seneca Lake	eDNA	SEN1	SEN	42.869	-76.941	0.0	0.0	23	–
7/25/19	Seneca Lake	eDNA	SEN2	SEN	42.869	-76.941	0.0	0.0	9	–
7/25/19	Seneca Lake	eDNA	SEN3	SEN	42.869	-76.941	0.0	0.0	3	–
7/25/19	Cayuga Lake	Blank	CAYBL	CAY	42.896	-76.75	0.0	0.0	3	–
7/25/19	Cayuga Lake	eDNA	CAY1	CAY	42.896	-76.75	269.2	1.8	49824	X
7/25/19	Cayuga Lake	eDNA	CAY2	CAY	42.896	-76.75	224.8	6.9	10958	X
7/25/19	Cayuga Lake	eDNA	CAY3	CAY	42.896	-76.75	127.4	4.1	11218	–
7/26/19	Cross Lake	Blank	CROBL	CRO	43.15	-76.483	0.0	0.0	3	–
7/26/19	Cross Lake	eDNA	CRO1	CRO	43.15	-76.483	314.3	1.8	15239	–
7/26/19	Cross Lake	eDNA	CRO2	CRO	43.15	-76.483	380.2	3.7	8	–
7/26/19	Cross Lake	eDNA	CRO3	CRO	43.15	-76.483	27.6	1.0	2131	–
7/26/19	Onondaga Lake	Blank	ONOBL	ONO	43.115	-76.241	0.0	0.0	3	–
7/26/19	Onondaga Lake	eDNA	ONO1	ONO	43.115	-76.241	1073.3	0.9	19398	–
7/26/19	Onondaga Lake	eDNA	ONO2	ONO	43.115	-76.241	2011.6	22.6	23	–
7/26/19	Onondaga Lake	eDNA	ONO3	ONO	43.115	-76.241	4746.1	65.6	122363	X
7/30/19	Lake Erie (East)	Blank	ERIEBL	ERIE	42.342	-79.595	0.0	0.0	3	–
7/30/19	Lake Erie (East)	eDNA	ERIE1	ERIE	42.342	-79.595	167.6	2.6	13721	–
7/30/19	Lake Erie (East)	eDNA	ERIE2	ERIE	42.342	-79.595	167.8	0.8	1882	–
7/30/19	Lake Erie (East)	eDNA	ERIE3	ERIE	42.342	-79.595	492.2	9.5	15907	–
7/31/19	Lake Erie (West)	Blank	ERIWBL	ERIW	42.08	-83.195	0.0	0.0	3	–
7/31/19	Lake Erie (West)	eDNA	ERIW1	ERIW	42.08	-83.195	96.8	1.0	277	–
7/31/19	Lake Erie (West)	eDNA	ERIW2	ERIW	42.08	-83.195	114.9	3.2	1459	–
7/31/19	Lake Erie (West)	eDNA	ERIW3	ERIW	42.08	-83.195	50.4	0.0	2924	–
7/31/19	Lake St Clair	Blank	LSCBL	LSC	42.564	-82.784	0.0	0.0	8	–
7/31/19	Lake St Clair	eDNA	LSC1	LSC	42.564	-82.784	685.2	6.4	66005	X
7/31/19	Lake St Clair	eDNA	LSC2	LSC	42.564	-82.784	1615.9	81.3	107722	X
7/31/19	Lake St Clair	eDNA	LSC3	LSC	42.564	-82.784	1072.7	10.8	16192	X
8/1/19	Lake Huron	Blank	LHSBL	LHS	43.43	-82.539	0.0	0.0	0	–
8/1/19	Lake Huron	eDNA	LHS1	LHS	43.43	-82.539	1116.9	9.3	19187	X
8/1/19	Lake Huron	eDNA	LHS2	LHS	43.43	-82.539	2331.2	15.7	15811	X
8/1/19	Lake Huron	eDNA	LHS3	LHS	43.43	-82.539	640.9	14.1	11805	X
8/2/19	Lake Michigan (east)	Blank	LMMBL	LMM	43.224	-86.339	0.0	0.0	0	–
8/2/19	Lake Michigan (east)	eDNA	LMM1	LMM	43.224	-86.339	358.1	9.6	34	–
8/2/19	Lake Michigan (east)	eDNA	LMM2	LMM	43.224	-86.339	1029.6	22.4	19	–

8/2/19	Lake Michigan (east)	eDNA	LMM3	LMM	43.224	-86.339	5118.8	72.2	137620	X
8/8/19	Lake Michigan (west)	Blank	LMKBL	LMK	42.579	-87.815	0.0	0.0	14	–
8/8/19	Lake Michigan (west)	eDNA	LMK1	LMK	42.579	-87.815	1956.7	4.6	19	–
8/8/19	Lake Michigan (west)	eDNA	LMK2	LMK	42.579	-87.815	6261.5	16.1	25	–
8/8/19	Lake Michigan (west)	eDNA	LMK3	LMK	42.579	-87.815	1686.3	6.0	23	–
9/18/19	Lake Erie (Buffalo)	Blank	BUFBL	BUF	42.846	-78.864	0.0	0.0		–
9/18/19	Lake Erie (Buffalo)	eDNA	BUF1	BUF	42.846	-78.864	322.2	5.1		–
9/18/19	Lake Erie (Buffalo)	eDNA	BUF2	BUF	42.846	-78.864	2609.7	9.3		–
9/18/19	Lake Erie (Buffalo)	eDNA	BUF3	BUF	42.846	-78.864	680.6	12.4		–
9/25/19	Oneida Lake	Blank	ONEBL	ONE	43.174	-75.93	0.0	0.0	0	–
9/25/19	Oneida Lake	eDNA	ONE1	ONE	43.174	-75.93	1524.9	4.9	80396	X
9/25/19	Oneida Lake	eDNA	ONE2	ONE	43.174	-75.93	1863.8	7.9	88127	X
9/25/19	Oneida Lake	eDNA	ONE3	ONE	43.174	-75.93	1567.9	4.1	17826	X
9/25/19	Lake Ontario (East)	Blank	OSWBL	OSW	43.463	-76.519	0.0	0.0	4	–
9/25/19	Lake Ontario (East)	eDNA	OSW1	OSW	43.463	-76.519	734.1	2.3	5297	–
9/25/19	Lake Ontario (East)	eDNA	OSW2	OSW	43.463	-76.519	584.2	0.0	18	–
9/25/19	Lake Ontario (East)	eDNA	OSW3	OSW	43.463	-76.519	6785.5	22.3	135672	X
9/28/19	Erie Canal	Blank	ECANBL	ECAN	43.119	-77.64	0.0	0.0	4	–
9/28/19	Erie Canal	eDNA	ECAN1	ECAN	43.119	-77.64	447.0	3.3	719	–
9/28/19	Erie Canal	eDNA	ECAN2	ECAN	43.119	-77.64	287.3	0.5	3271	–
9/28/19	Erie Canal	eDNA	ECAN3	ECAN	43.119	-77.64	399.7	1.7	10622	–
9/28/19	Lake Ontario (West)	Blank	ROCBL	ROC	43.27	-77.626	0.0	0.0	4	–
9/28/19	Lake Ontario (West)	eDNA	ROC1	ROC	43.27	-77.626	542.7	0.0	60	–
9/28/19	Lake Ontario (West)	eDNA	ROC2	ROC	43.27	-77.626	821.1	5.5	6	–
9/28/19	Lake Ontario (West)	eDNA	ROC3	ROC	43.27	-77.626	1022.7	3.1	1818	–

Tissues

Date	Body of water	Type	Sample_ID	Site_ID	Latitude	Longitude	Number of individuals
7/30/19	Lake Erie (East)	Tissue	ERIE1–ERIE13	ERIE	42.342	-79.595	13
7/31/19	Lake Erie (West)	Tissue	ERIW1–ERIW9	ERIW	42.080	-83.195	9
7/31/19	Lake St Clair	Tissue	LSC1–LSC14	LSC	42.564	-82.784	14
8/1/19	Lake Huron	Tissue	LHS1–LHS28	LHS	43.430	-82.539	28
8/2/19	Lake Michigan (east)	Tissue	LMM1–LMM28	LMM	43.224	-86.339	28
8/8/19	Lake Michigan (west)	Tissue	LMK1–LMK28	LMK	42.579	-87.815	28
9/7/19	Cross Lake	Tissue	CRO1–CRO18	CRO	43.150	-76.483	18
9/11/19	Cayuga Lake	Tissue	CAY1–CAY28	CAY	42.896	-76.750	15
9/11/19	Onondaga Lake	Tissue	ONO1–ONO28	ONO	43.115	-76.241	28
9/28/19	Erie Canal	Tissue	ECAN1–ECAN20	ECAN	43.119	-77.640	20
9/28/19	Lake Ontario (West)	Tissue	ROC1–ROC28	ROC	43.270	-77.626	28
9/25/19	Oneida Lake	Tissue	ONE1–ONE28	ONE	43.174	-75.930	28
9/25/19	Lake Ontario (East)	Tissue	OSW1–OSW28	OSW	43.463	-76.519	28

Table S4.2. The qPCR primer/probe assays used in this study. The stop codon inserted into the COI gBlock fragment is indicated in bold.

Marker	Assay	Sequence 5' – 3'
mtDNA	GobyCOI-F2*d*	CTTCTGGCCTCCTCTGGTGTTG
	GobyCOI-R2d	CCCTAGAATTGAGGAAATGCCGG
	GobyCOI-Pr	6-FAM-CAGGCAACTTGGCACATGCAG-BHQ
	GobyCOI-gBlock	TGATTGCTCCCTCCTTCTTTCTTACTACTTCTGGCCTCCTCTGGTGTTG AAGCAGGGGCA**AGA**ACCGGGTGGACAGTTTACCCTCCCCTGGCAGGC AACTTGGCACATGCAGGAGCATCCGTCGACTTGACAATCTTCTCCCTT CACCTGGCCGGCATTTCCTCAATTCTAGGGGCTATTAATTTTATTA
nuDNA	*Nmel*505-F	CAACAGCACCCAATACCAGATC
	*Nmel*505-R	GAAGGCTGACAGTGGGCATT
	*Nmel*505-Pr	6-FAM-ACACTGAGCTGGATGC-BHQ
	*Nmel*505-gBlock	TCTACCTATTTATCTAATGACCACTCAACAGCACCCAATACCAGATCT ACACTGAGCTGGATGCAAATGCCCACTGTCAGCCTTCTGTCTTTGTCG CTTTAGGTCCGTACCTGACAAAATATTCA

Table S4.3. Summary information for 28 microsatellite loci used in the study*, including the primer name, repeat motif, forward and reverse primer sequences, multiplex number (*MP*), number of alleles (N_A), minimum and maximum allele length (L_{MIN}, L_{MAX}), and average observed heterozygosity (H_O) in genotyped Round Goby tissues. Denoted loci (†) were excluded from analysis due to poor amplification.

Primer	Tetramer repeat	Forward sequence (5'-3')	Reverse sequence (5'-3')	MP	N_A	L_{MIN}	L_{MAX}	H_o
Nmel140	AGAT	ATGTGCTGGATTGTCTCATTGG	AGCTTGTTAATATGGCACGGTG	4	11	231	435	0.516
Nmel262	ATCC	CGACAGCTGTAGTTCAACTCAC	AAACGGACAAGTGGAAGAAG	1	7	236	270	0.433
Nmel299	AGAT	TTATGCTCACTACTGCCACTTC	AGGAAACAAGGAAGGGACAAAC	1	12	230	367	0.566
Nmel32	AGAT	CTCCAAACCTTCTTACTGCTGC	AACCCTCACCTGAATCAAGCAC	4	4	276	385	0.330
Nmel344	AGAT	GCTATAATGTTGTGTGCAGTGC	ACACTGCCTTAATGATTCAAAGG	5	9	229	354	0.074
Nmel351	AGAT	TTTGGTCACTGGAATTACAATACG	CATATTGGCTTCAGTAGATGTTTG	1	12	228	256	0.499
Nmel363	AAGT	TCAAAGAGACCTAAGCCAGTCC	CTGTTTGGCTCGGACAACATG	1	9	259	299	0.470
Nmel403	AGAT	GTCTTGCTGCATAAACCAAAGC	GGACCGATCTGTATATAGGAGCC	5	10	259	301	0.764
Nmel405	AGAT	TCACAGCTAAGACCAACAACC	AATGGACTAACAGTTCGCTAC	5	15	235	339	0.379
Nmel411	ATCC	GGAAACAAGGTCGCAGGTAATG	CAGCAACGGACAAGTGGAAG	2	8	232	284	0.487
Nmel422	ATCC	CTCCCATCTCCCATCCAAGTAC	TACAAAGCGCGATATCAGTTGG	2	3	244	316	0.488
Nmel505	AGAT	CCCACTTCAATGTTCTGCACTC	GCTTTCACCTCATTCATACAGC	5	5	239	263	0.029
Nmel549	AGAT	TGCATGATATAGGACCTTTAAAGC	TGGACTGTAAGCCATAATCTTC	6	11	229	281	0.492
Nmel625	ATCC	CAGACAAAGAGCGGTTCAGAAG	TGATGGTGGGTGAAGTACATGG	6	10	229	323	0.533
Nmel660†	ATCC	CTGGACACAAGCTGCAAGTG	CTATGTGATTTGGGCAGCTCC	2	2	230	278	0.354
Nmel729	ACAG	ACAAGTGCATTAGTGTCATGGG	AAGGCCTAACAAACAGCAGATG	2	5	237	242	0.004
Nmel746	AAAG	GTTTATGCCCTGGACATCTGC	GCTCGAGTCTTAAATGCAGATTG	6	6	237	277	0.440
Nmel810	ATCC	AACACAGCGTCAAATCTCTCAG	ATTACTAGTAGGTGACGGGTCG	7	7	230	274	0.386
Nmel815	AGAT	GGGTGTATTGTCTAGTCTCTGG	CAGGCTCATGTTAAGGGTTCAG	6	12	240	295	0.277
Nmel821	ACAT	CCACCAATCATGTTACACAGGG	AGTTCAGAGGCAGCCAGTTG	7	7	263	321	0.576
Nmel89	ACAT	TCTACTTCAGTTGGCTATGATTC	TAAGAGCAGGCTTAAGACCCAC	7	6	230	266	0.396
Nmel914	AGAT	TTGTGGACAAGGGCTGAACAG	CTATACGCACAGCTTCCTCACC	3	6	241	295	0.408
Nmel990	AGAT	AAATGCTGTTATACTGAGGCGG	CTCAGGGCTCCAACTACTCC	3	12	234	302	0.458
Nmel994	AGAT	CTCTCTGACATGCTCCAAGGAC	TTCTCAATAACTTTATCTCGGACC	3	11	237	265	0.530
Nmel1132	AGAT	TCCTGGATGAACACTACAAGGC	AGGACGTTTCGCTTTGATCTTC	7	12	247	271	0.645
Nmel1462	AAAG	CGATACCCAATATAGCGGACGG	AGTCAGAAGGAAACACATGCAG	5	8	232	289	0.361
Nmel1486	ATCC	TTCCCATTGTAACAGCAGAGAG	AATGTGATTTGGGCAGCTCG	6	14	249	360	0.686
Nmel1566	AGAT	ATTTCGGATCCTCGTATCTGAC	AATGATATGGGAAATGCGAGCG	3	19	231	435	0.516

* Loci were developed in Andres et al. (2021).

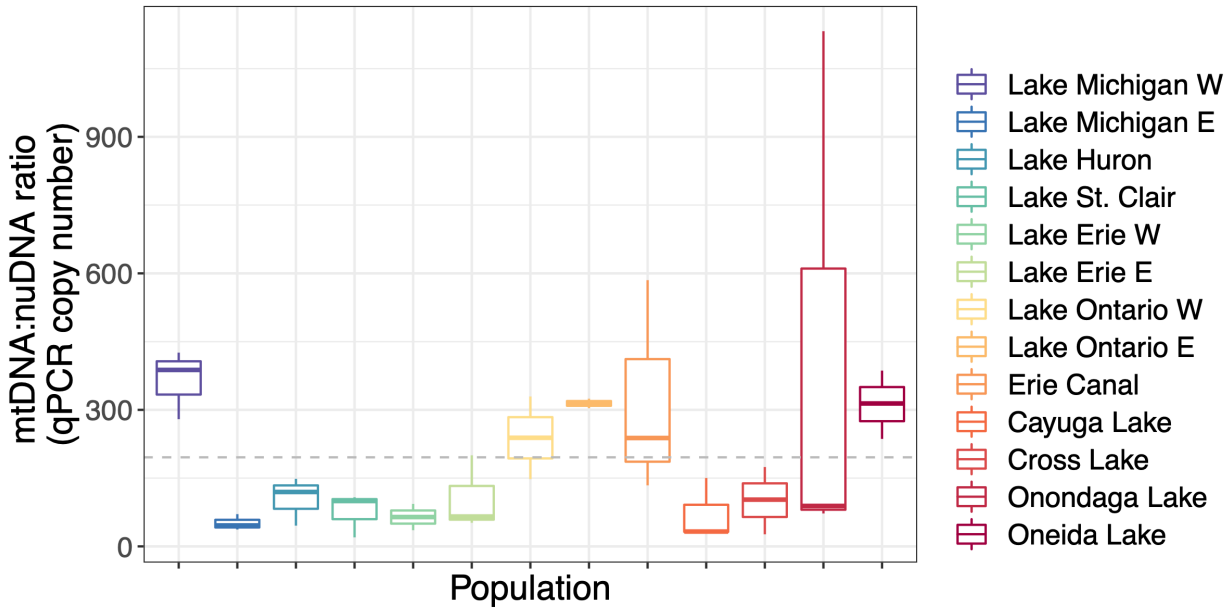

Figure S4.1. Boxplot showing the mean ratio of mtDNA to nuDNA across eDNA samples per population (horizontal line), first and third quartiles (hinges of the box), and 1.5 * the interquartile range (upper/lower whiskers). The mean ratio (196:1) is indicated with the gray dashed line.

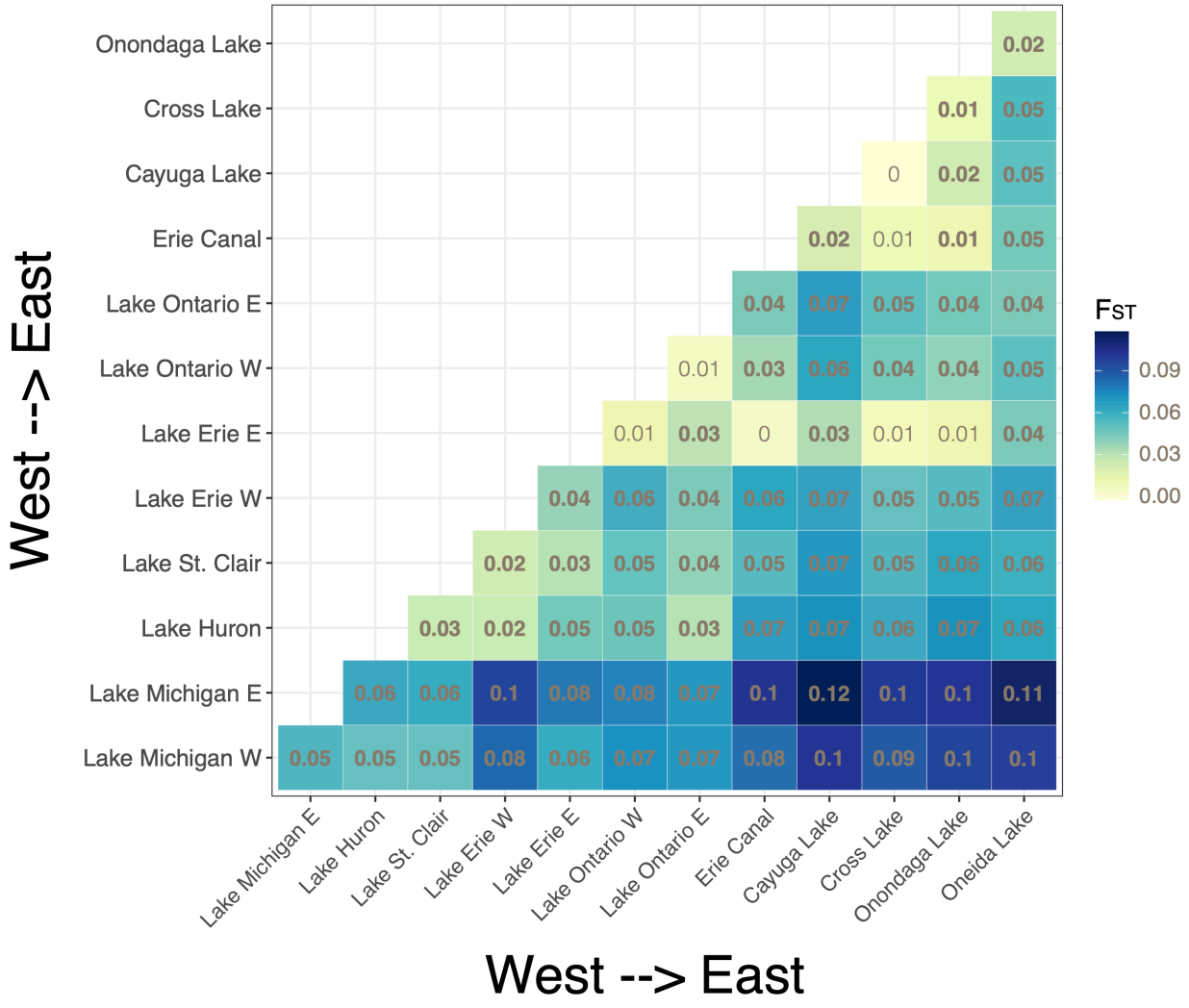

Figure S4.2. Pairwise F_{ST} comparisons based on tissue-based genetic analysis 27 microsatellite loci for Round Goby populations in North America. Each column and row represent a sampled population ordered along a longitudinal gradient. Significant pairwise comparisons (95% confidence intervals do not overlap with 0) are shown in bold.

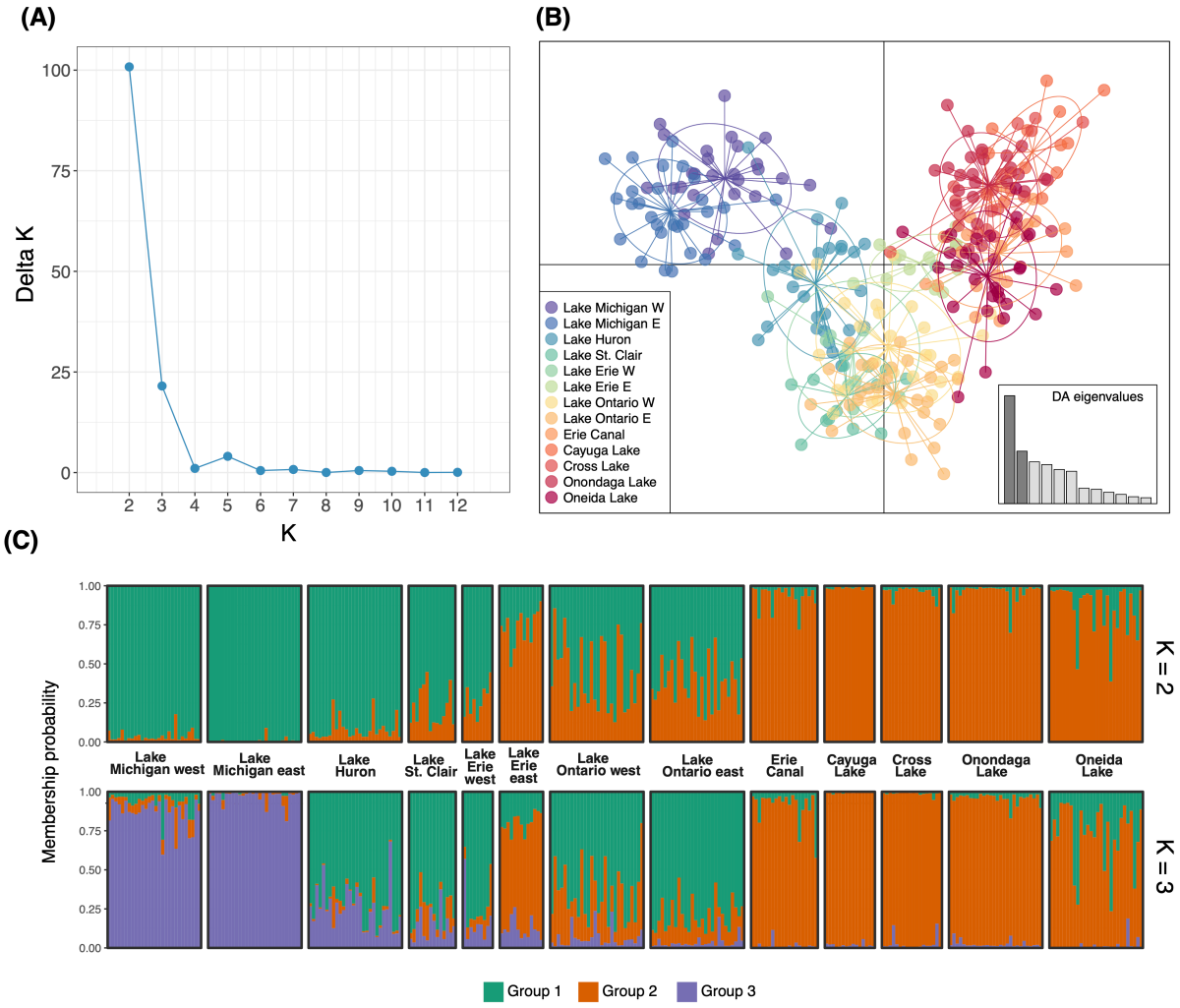

Figure S4.3. Spatial analysis of Round Goby population structure based on Round Goby tissues. (A) Delta-K statistic estimated for each value of K based on the Evanno method. The most probable value of K was determined to be K = 2; (B) Discriminant analysis of principle components (DAPC) scatter plot of 299 genotyped Round Gobies. Axes represent discriminant functions 1 and 2 and points represent individual Round Goby genotypes. The inset bar chart indicates the eigenvalues for each discriminant function; (C) Estimated assignment probabilities of the 299 Round Gobies at K = 2 and K = 3. The group membership

of each individual is represented by a single vertical line with the proportion of each segment indicating the assignment probability to each group.

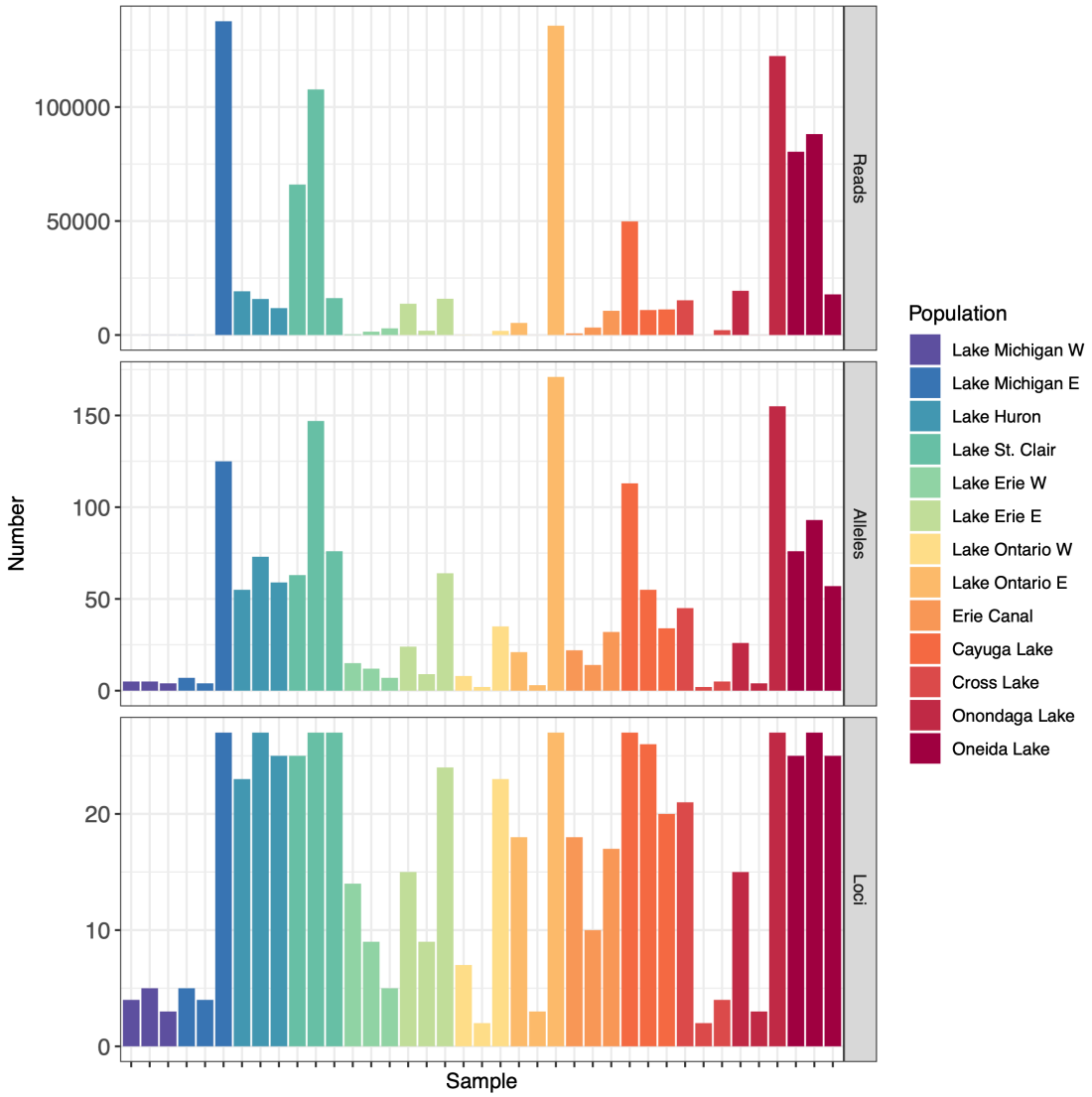

Figure S4.4. Read depth, number of alleles, and number of loci detected in each eDNA sample. Triplicate samples are colored per site.

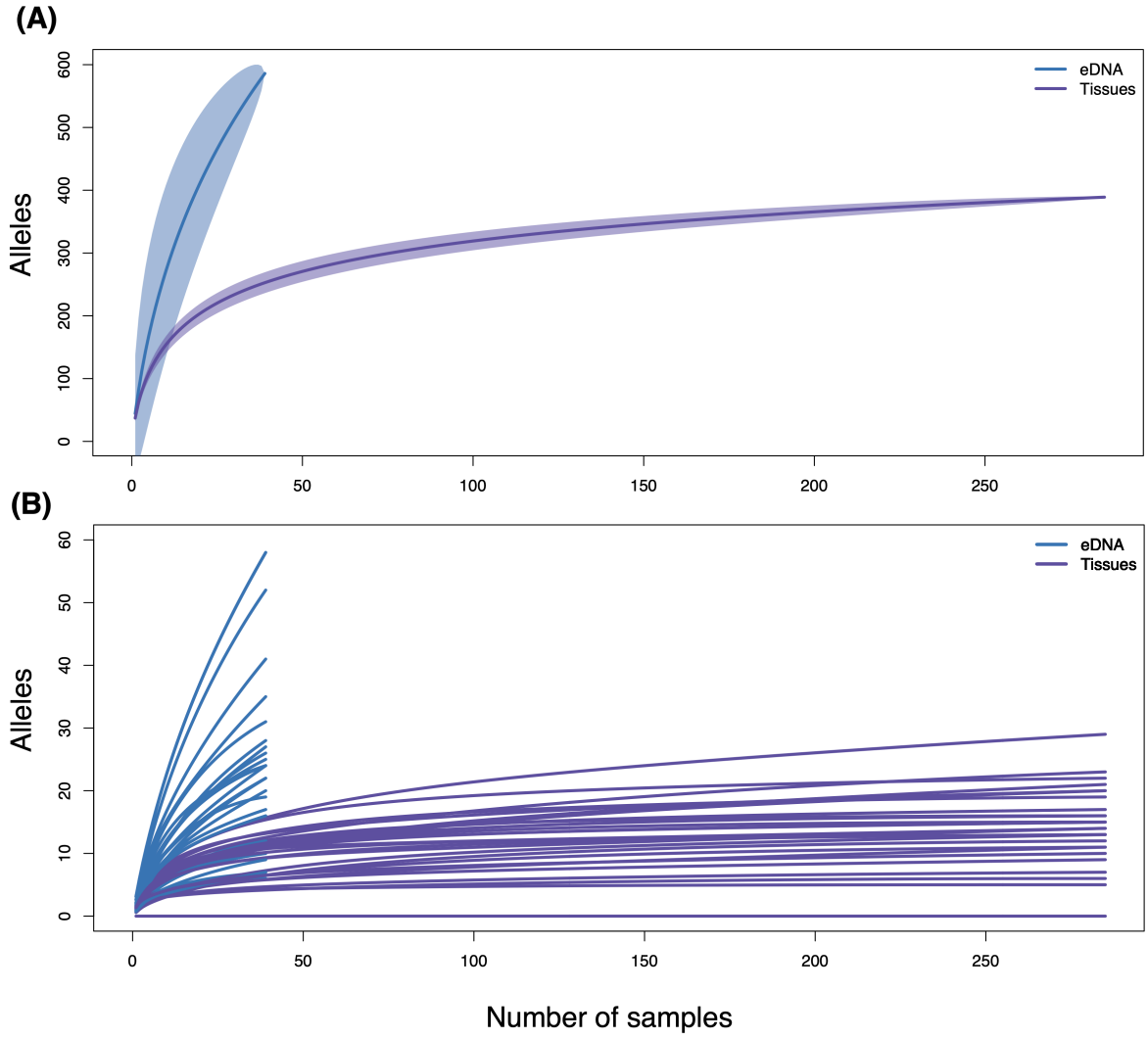

Figure S4.5. Accumulation curves showing the number of microsatellite alleles expected in a given number of Round Goby tissue samples and eDNA samples for (A) all loci combined and (B) each locus plotted separately.

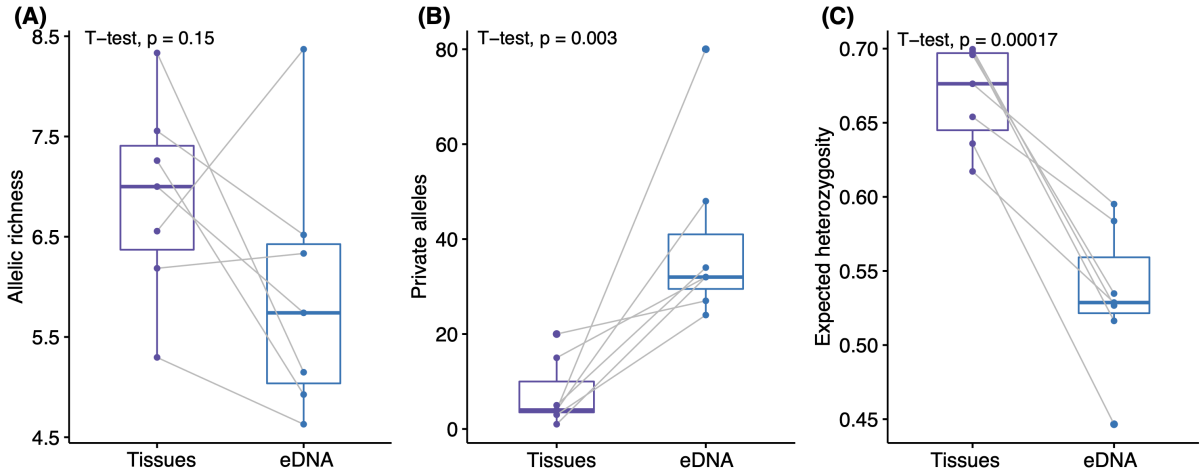

Figure S4.6. Summary statistics for population-level genetic data collected from genotyped Round Goby tissues and eDNA samples, including (A) average allelic richness across 27 microsatellite alleles, (B) number of private alleles, and (C) expected heterozygosity. Population estimates of each metric based on tissues and eDNA are connected with gray lines.

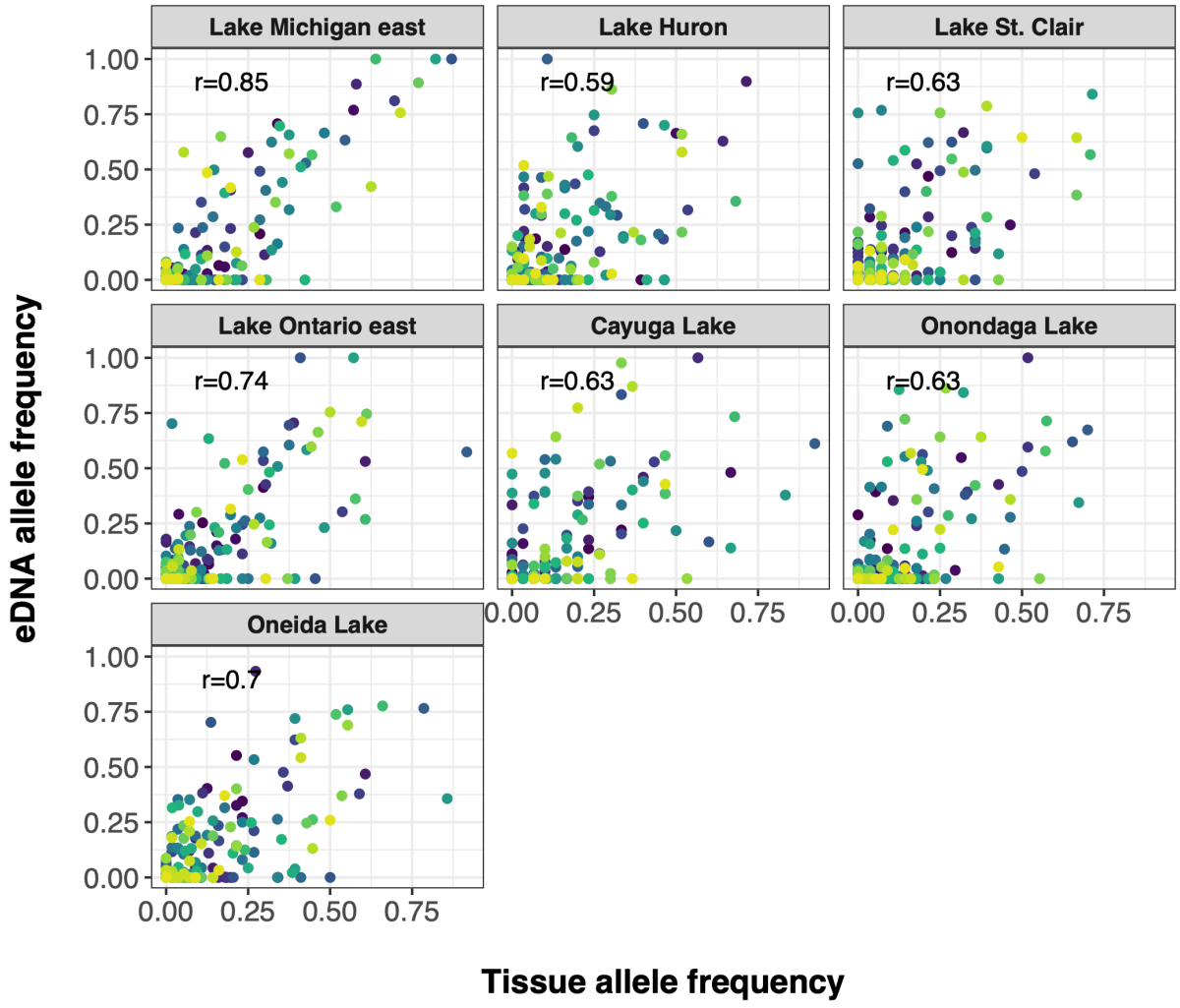

Figure S4.7. Correlation of allele frequencies per site. Each point represents the frequency of an allele in genotyped tissues and eDNA sequence reads, with different loci denoted in different colors.

APPENDIX 5

Supplementary Information for Chapter 5:

Estimating Round Goby abundance using advanced survey methods: a comparison of AUV imaging and eDNA assessments

Figure S5.1. Images of sampling equipment used in this study.

Figure S5.2. Sampling map and mitochondrial control region haplotypes for Round Goby tissues.

Figure S5.3. Bioinformatic processing information for d-loop sequences.

Supplementary methods S5.1. eDNA extraction protocol for Envirochek capsule filters.

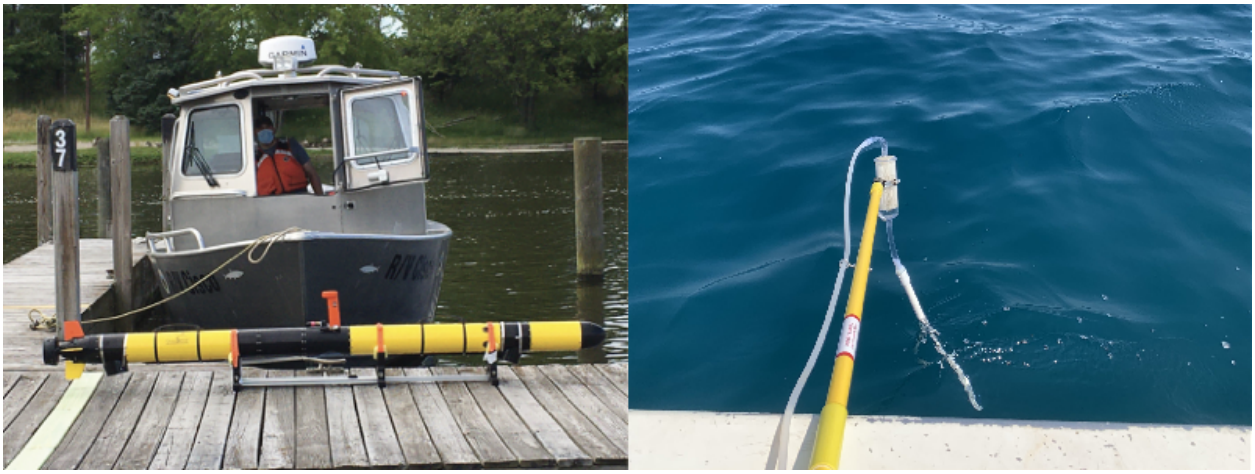

Figure S5.1. Images of sampling equipment used in this study. The autonomous underwater vehicle (AUV) was used to obtain benthic images of Round Gobies (left) and the Smith-Root eDNA Samper fitted with an Envirochek HV filter was used to collect large-volume eDNA samples (right).

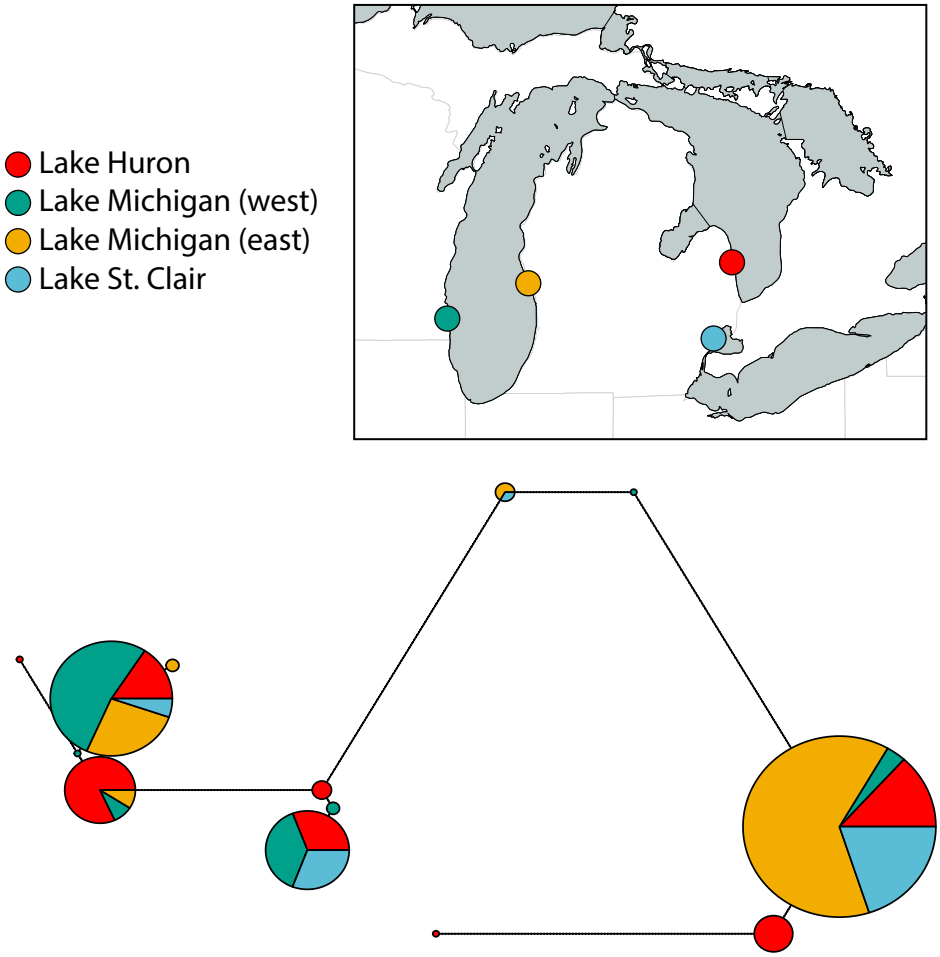

Figure S5.2. Sampling map and mitochondrial control region haplotypes for Round Goby tissues. Haplotype networks were constructed using the most common haplotype from each sampled individual. Each circle represents a unique haplotype, with the diameter of the circle representing the relative frequency of the haplotype across all individuals and the color slices of the circle representing the relative frequency of the haplotype across the sites where that haplotype was detected. The length of the line between haplotypes represents the number of mutations separating the haplotypes. Haplotype networks were generated using `pegas` (Paradis, 2010) in R.

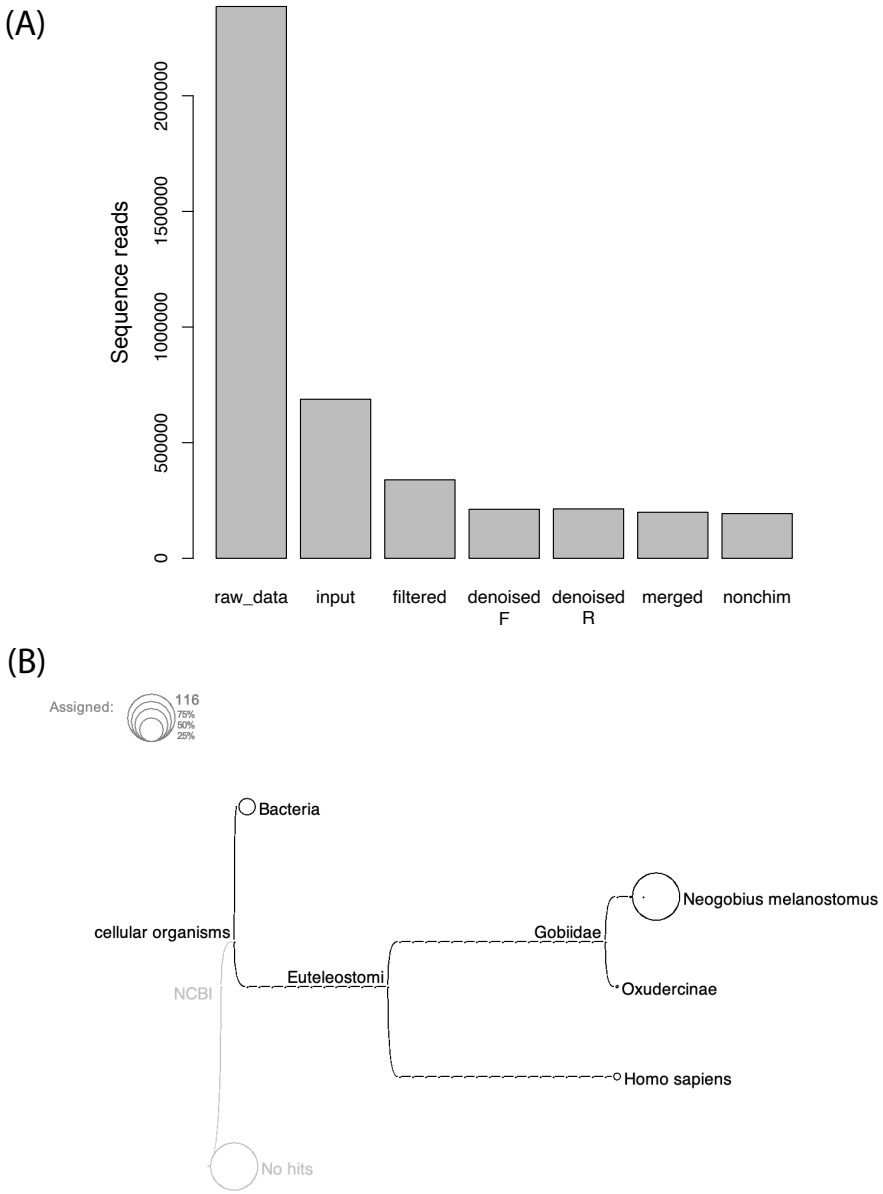

Figure S5.3. (A) Number of sequence reads at sequential bioinformatic processing stages, including raw reads, reads input to DADA2 following Trimmomatic, and stages of DADA2 including quality filtering, denoising, merging, and chimera removal. (B) Phylogenetic tree showing the taxonomic assignment of Illumina sequences targeting the Round Goby d-loop in

eDNA samples. Although many ASVs match to Round Goby (*Neogobius melanostomus*), many ASVs match to bacteria or do not return a BLASTn hit with similarity with an E-value exceeding $1E^{-5}$. The size of the circle represents the number of ASVs that matched each taxonomic group.

Supplementary methods S5.1. eDNA extraction protocol for Envirochek capsule filters.

1. Carefully wipe the outer surfaces of all capsule filters with 5% bleach using clean paper towels. Dry and wipe with 70% Ethanol.
2. Incubate filters at 56º C for 2 hours.
3. Agitate each filter for 5 min (manually or in a shaker). Uncap one end of the capsule filter and pour 40 mL of the sample into a 50 mL conical tube.
4. Complete 3 sequential washes with 10 mL clean preservation buffer each, totalling 30 mL. Pour the wash contents into a new 50 mL conical tube. Mix the contents of the tubes by pouring the buffer back and forth between them several times. The final volume in each tube should be 35 mL (70 mL total for each sample).
5. Pour the contents of each tube into a new 85 mL centrifuge tube.
6. Add 3.5 mL 3M Sodium Acetate (pH 5.5) and 35 mL room temperature isopropanol to each centrifuge tube. Cap and mix by inverting, then put in the -20 for at least an hour (up to overnight).
7. Centrifuge the samples at 5ºC at 10,000 g for 15 minutes. Decant supernatant, being careful not to disturb the pellet.
8. Add 10 mL 75% EtOH and centrifuge at 3,000 g for 2 min. Decant supernatant, being careful not to disturb the pellet. Remove any remaining EtOH and let the pellet air dry for 5 min.
9. Add 180 µL ATL and 20 µL ProK to each tube, mixing thoroughly after addition. Add mixture from both centrifuge tubes to the same 2 mL tube.
10. Close tube, seal with parafilm, and Vortex for 15 s.
11. Incubate while rotating at 56℃ for at least 2 hours (up to overnight).
12. Vortex each sample for 15 s and spin down for 2 s to seed out excess debris.
13. Add Buffer AL and ice-cold 100% EtOH to each tube in equal volumes (1:1:1: sample:Buffer AL:EtOH; 400 µL each). Vortex immediately after addition.
14. Add ~650 µL of the mixture to a DNeasy spin column. Centrifuge at 6,000 g (8,000 rpm) for 1 minute. Discard flow through. Place the column in a new 2 mL collection tube and repeat.
15. Add 500 µL AW1 to each sample. Centrifuge at 6,000 g (8,000 rpm) for 1 minute. Discard flow through and place the column in a new 2 mL collection tube.
16. Add 500 µL AW2 to each sample. Centrifuge at 20,000 g (14,000 rpm) for 3 minutes. Discard flow through and place the column in a new 2 mL collection tube.
17. Centrifuge 1 min at 17,000 g (13,000 rpm) to remove any remaining liquid. Discard flow through and place the column in a new 2 mL collection tube.
18. Add 50 µL 70°C Buffer AE (pH 8.0) to the membrane and incubate at RT for 10 min.
19. Centrifuge for 1 min at 6,000 g (8,000 rpm).
20. Repeat elution step by adding another 50 µL 70°C Buffer AE (pH 8.0) to the membrane. Incubate at RT for 10 min.
21. Centrifuge for 1 min at 6,000 g (8,000 rpm).
22. Discard the spin column. Transfer eluted DNA to a new 1.5 mL DNA LoBind tube.